RANDOM EFFECTS MODELS FOR NOMINAL AND ORDINAL DATA

By

JONATHAN SETH HARTZEL

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

1999

To my family, Tracy, Riley, and Kendi.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Dr. Alan Agresti for serving as my dissertation advisor and for providing me the opportunity to work with him as a research assistant. During the past two years as his research assistant, I have gained invaluable experience in both conducting statistical research and producing scholarly writing. His obvious interest in both my present research and my future as a professional has been a constant encouragement to me. I would also like to thank Drs. Malay Ghosh, James Hobert, Ramon Littell, and Gary Miller for serving on my committee. In addition, I thank Dr. Jane Pendergast for her time on my committee before leaving the University of Florida.

I thank the Lord for giving me the strength and the perseverance to reach this point, and for giving my wife, Tracy, the patience and understanding through it all. Her countless sacrifices for me and unwavering confidence in me were the cornerstone for my success. I am forever grateful.

I thank my parents for their love, their continual support of my educational pursuits, and their constant prayers. In addition, I thank my brother and sister for the many cards and pictures of the kids, and for telling them all about Uncle Jonathan and Aunt Tracy while we have been away. I also want to thank my mother- and father-in-law for their immediate love and support from the first time I met them. Their visits to Gainesville were much appreciated. Finally, I thank Riley and Kendi for always being excited to see me when I came home from school.

TABLE OF CONTENTS

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

RANDOM EFFECTS MODELS FOR NOMINAL AND ORDINAL DATA

By

Jonathan Seth Hartzel

December 1999

Chairman: Alan G. Agresti
Major Department: Statistics

Models for nominal and ordinal response data are important in many areas of research. In medical studies, patients are often evaluated on an ordinal or graded scale. Nominal data, such as types of services used at a hospital, are frequent in the field of health care. It is often the case that such data are nested within clusters or repeatedly assessed over time. In this dissertation we propose random effects models for analyzing longitudinal or clustered nominal and ordinal response data. Specifically, we present a general multinomial logit random effects model that we motivate within the framework of a multivariate generalized linear mixed model. As special cases of the proposed model, we consider models based on the cumulative logit, adjacent-category logit, and continuation-ratio logit link functions for analyzing ordinal response data, and the baseline-category logit link function for nominal data.

For the proposed multinomial random effects models, we consider both parametric and nonparametric assumptions for the distribution of the random effects. In the parametric approach, we assume that the random effects follow a multivariate normal distribution. We consider direct maximization of the marginal likelihood using adaptive Gauss-Hermite quadrature, as well as indirect maximization using an automated Monte Carlo expectation-maximization (EM) algorithm. In addition, we propose a pseudo-likelihood approach for obtaining approximate maximum likelihood estimates. In the nonparametric approach, we assume that the random effects follow an unspecified discrete distribution. We propose an EM algorithm for obtaining nonparametric maximum likelihood estimates of the model parameters and the discrete distribution. Using simulation, we compare the performance of the parametric and nonparametric approaches when the random effects distribution is misspecified.

We also examine the use of the proposed models for modeling ordinal multi-center clinical trial data. We consider random effects models that allow for a common association across all centers, as well as heterogeneous associations. We also propose Laplace and adaptive Gauss-Hermite quadrature approximated score tests for testing that a common association parameter holds in the heterogeneous random effects model. We show that the latter test performs poorly for small to moderate numbers of centers.

CHAPTER 1
METHODS FOR MODELING LONGITUDINAL AND CLUSTERED DATA

1.1 Introduction

In many areas of research data are collected on the same experimental unit over time or under multiple conditions. Longitudinal studies in which measurements are taken on the same subject over time occur frequently in medical and biological research. Repeated measures data are common in the social and behavioral sciences where each treatment or condition is applied to each subject in an effort to control between subject variability. Data may also be collected in groups or clusters. Surveys and observational studies of populations that have a natural hierarchical structure, such as students nested within classrooms, lead to the collection of clustered data. In multi-center clinical trials, multiple treatments are compared at different sites resulting in data that are clustered at the center level. Regardless of the sampling mechanism, observations taken on the same subject or from the same cluster are often correlated. Use of traditional linear or generalized linear models for modeling and inference in this setting would be incorrect as the assumption of independent observations is violated. Special modeling techniques are needed that can account for the correlation within subjects and clusters.

One technique for incorporating correlation that has become increasingly popular involves the use of random effects. Models that incorporate random effects appear throughout the literature under a variety of names such as random effects models (Laird and Ware 1982; Stiratelli et al. 1984), mixed effects models (Harville and Mee 1984), variance component models (Harville 1977), and random coefficient models (Longford 1993). The basic idea underlying a random effects model is that heterogeneity exists across subjects in all or some subset of their regression coefficients. This variability may be attributable to, for example, unmeasured covariates or imperfect measurement of measured covariates. It is assumed that the heterogeneity can be represented by a probability distribution. To account for the variability, an unobserved random variable from the probability distribution is incorporated additively in the model. Since observations from the same subject share the same unobserved realization, correlation is induced between the observations within the subject. Random effects models are defined conditionally upon the random effects. Estimates of the fixed and random parameters are obtained by maximizing the marginal likelihood, which requires integrating the joint likelihood over the random effects.

Statistical methods for correlated data, and software packages for implementing these methods, are readily available when the data consist of correlated normal responses. In this case when the random effects distribution is also assumed to be normal, estimation of both the fixed and random effects is straightforward as the marginal likelihood can be written in closed form. A detailed history of the linear mixed model (LMM) can be found in Searle et al. (1992). In the past twenty years there has been a considerable amount of research in the area of correlated data where the response is non-normal. In particular much attention has focused on correlated Poisson, Bernoulli, and binomial response data. A survey article by Pendergast et al. (1996) listed well over one hundred references for the analysis of correlated binary data alone! Random effects models in this context are often derived as extensions of generalized linear models (GLMs) (Nelder and Wedderburn 1972; McCullagh and Nelder 1989) and referred to as generalized linear mixed models (GLMMs) (Gilmour et al. 1985). In contrast to linear mixed models, the assumption of a normal random effects distribution leads to an intractable marginal likelihood. Thus a majority of the recent literature in this area has focused on methods for approximating the marginal likelihood (Zeger and Karim 1991; McCulloch 1997; Booth and Hobert 1999).

In comparison, there has been relatively little research for nominal and ordinal response data. A nominal response is a categorical variable with unordered levels, whereas an ordinal response has ordered levels. Models for nominal and ordinal response variables assume that the counts within each category of the response follow a multinomial distribution for each combination of the covariates.

The majority of the work for multinomial response data has focused on random effects models for ordinal responses, with the random effects assumed to be normal. As in the binary case, this leads to an intractable marginal likelihood. Harville and Mee (1984) were among the first to have proposed a random effects model for ordinal data, fitting the model using a best linear unbiased prediction procedure (Henderson 1975). They utilized a first-order Taylor series approximation for evaluation of the intractable integrals. Simpler models that allowed for only single random effects and utilized numerical integration were proposed later by Jansen (1990) and Ezzet and Whitehead (1991). In recent work by Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996), general random effects regression models for ordinal responses were proposed along with a variety of estimation procedures. All of these previous models have used either the cumulative logit or cumulative probit links. In contrast, Ten Have and Uttal (1994) and Ten Have (1996) proposed ordinal random effects models based on the continuation-ratio logit link and the cumulative complementary log-log link, respectively. In the latter model, the random effects distribution was assumed to be log-gamma, which resulted in a closed form marginal likelihood.

For nominal response data, Fahrmeir and Tutz (1994, p. 231) proposed a random effects model based on the baseline-category logit model as did Hedeker (2000). Greater attention has been given to a special case of this model in the psychometric literature, however. In psychometric research, a popular class of qualitative response models is the Rasch family of models (Rasch 1961). Such models can be considered as baseline-category logit models. Adams and Wilson (1996) considered a baseline-category logit Rasch model that allowed for shifted thresholds. This work was then extended by Adams et al. (1997) to allow for varying thresholds. We will consider these models in greater detail in Chapter 3.

The focus of this dissertation will be on random effects models for nominal and ordinal data. The models considered will be motivated as extensions to multivariate generalized linear models (MGLMs). Careful attention will be given to estimation methods when the random effects are assumed to be normally distributed. We also propose a model in which this assumption is relaxed, resulting in a nonparametric random effects model. Applications of these models to multi-center clinical trial data will be examined in detail. We begin by reviewing in greater detail some of the work that has been discussed above. In Section 1.2 we will briefly consider models for normal responses. Greater attention will be given to the non-normal response case in Section 1.3. There we will delineate between subject-specific models and population-averaged models. In the final section of this chapter an outline of the remainder of the dissertation will be given.

1.2 Normal Response Data

As noted before, there has been an extensive amount of research on LMMs for the analysis of longitudinal data (see e.g., Jones 1993; Lindsey 1993; Diggle et al. 1994). We review one such model, originally proposed by Harville (1977) and further developed by Laird and Ware (1982), that has been influential in the development of models for non-normal responses. Let $\mathbf{y}_i$, $i = 1, \cdots, n$, be the response vector for the ith subject. Harville (1977) proposed the linear mixed model

$$\mathbf{y}_i = Z_i\boldsymbol{\beta} + W_i\mathbf{u}_i + \boldsymbol{\epsilon}_i \tag{1.1}$$

where $\boldsymbol{\beta}$ is a vector of fixed effect parameters, $\mathbf{u}_i$ is a vector of random effects, and Z_i and W_i are corresponding design matrices. In addition, the vectors $\boldsymbol{\epsilon}_i$, $i = 1, \cdots, n$, are independent and distributed as $N(\mathbf{0}, R_i)$. The separation of the fixed and random components in model (1.1) is now the standard for expressing LMMs and GLMMs. The model formulation is conditional on $\mathbf{u}_i$ and the definition is completed by assuming that the $\mathbf{u}_i$ are distributed as $N(\mathbf{0}, Q)$, independently of each other and of the $\boldsymbol{\epsilon}_i$. Thus marginally

$$\mathbf{y}_i = Z_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i^*, \tag{1.2}$$

where $\boldsymbol{\epsilon}_i^* \sim N(\mathbf{0}, V_i)$ and $V_i = R_i + W_iQW_i'$. The covariance matrices, R_i and Q, are typically functions of unknown parameters, and allow for modeling of both within subject and between subject associations.

Let $\boldsymbol{\theta}$ be the vector of parameters for the covariance matrix V_i. If $\boldsymbol{\theta}$ is known, then the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$, calculated from the model (1.2), is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} Z_i'V_i^{-1}Z_i\right)^{-1} \sum_{i=1}^{n} Z_i'V_i^{-1}\mathbf{y}_i, \tag{1.3}$$

and is equal to the weighted-least-squares solution. Harville (1977), as well as others, showed that the MLE of $\boldsymbol{\beta}$ is also a best linear unbiased estimator (BLUE). Due to the assumptions of normality and the independence of $\mathbf{u}_i$ and $\boldsymbol{\epsilon}_i$, the predicted posterior means (modes) of $\mathbf{u}_i$ are easily shown to be

$$\hat{\mathbf{u}}_i = QW_i'V_i^{-1}(\mathbf{y}_i - Z_i\hat{\boldsymbol{\beta}}). \tag{1.4}$$

In practice the components $\boldsymbol{\theta}$ are not known and must be estimated. Harville (1977) considered both maximum and restricted maximum likelihood estimation (REML) of $\boldsymbol{\theta}$. The ML estimates of $\boldsymbol{\theta}$ are calculated by maximizing the marginal log-likelihood, based on the marginal model (1.2), with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. The REML estimates can be derived using two unrelated approaches (Laird and Ware 1982). One method of obtaining the REML estimates is by maximizing the likelihood of $\boldsymbol{\theta}$ based on a linearly transformed set of data $\mathbf{y}^* = A\mathbf{y}$ such that the distribution of $\mathbf{y}^*$ does not depend on $\boldsymbol{\beta}$. The transformed data $\mathbf{y}^*$ can be obtained by choosing A to be the matrix that converts $\mathbf{y}$ to the ordinary least-squares residuals. The REML estimates can also be derived from a Bayesian perspective. This is achieved by considering $\boldsymbol{\beta}$ in model (1.1) as a random variable having a vague or totally flat prior distribution such that the prior density of $\boldsymbol{\beta}$ is a constant. For example, let

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \Gamma), \quad \text{with} \quad \Gamma \to \infty,$$

where the elements of Γ go to infinity. The REML estimates are found by maximizing the limiting (as $\Gamma^{-1} \to 0$) marginal log-likelihood of $\boldsymbol{\theta}$ given $\mathbf{y}$. Iterative methods are needed to calculate either the ML or REML estimates of $\boldsymbol{\theta}$ since the likelihoods are nonlinear functions of $\boldsymbol{\theta}$. Harville (1977) proposed Newton-Raphson and scoring algorithms to estimate $\boldsymbol{\beta}$, $(\mathbf{u}_1, \cdots, \mathbf{u}_n)$, and $\boldsymbol{\theta}$. A number of expectation-maximization (EM) algorithms (Dempster et al. 1977) for jointly estimating $(\boldsymbol{\beta}, \mathbf{u}_1, \cdots, \mathbf{u}_n, \boldsymbol{\theta})$ have been proposed as well (Laird and Ware 1982; Jones 1993).

1.3 Non-Normal Response Data

We now review models for longitudinal and clustered data when the response is non-normal. The majority of these models can be viewed as extensions to generalized linear models (Nelder and Wedderburn 1972; McCullagh and Nelder 1989) and so we begin by defining GLMs. Let y_i, $i = 1, \cdots, n$, be the response for the ith subject and $\mathbf{z}_i$ the corresponding design vector. The definition of a GLM consists of a distributional assumption and a structural assumption.

1. Distributional Assumption:

   Conditional on $\mathbf{z}_i$, the y_i are independent and have a distribution in the exponential family with $E(y_i \mid \mathbf{z}_i) = \mu_i$.

2. Structural Assumption:

   The expectation μ_i is related to the linear predictor $\eta_i = \mathbf{z}_i'\boldsymbol{\beta}$ by the link function

$g(\cdot)$ or, equivalently, by the response function $h(\cdot) = g^{-1}(\cdot)$ such that

$$\eta_i = g(\mu_i) \quad \text{or} \quad \mu_i = h(\eta_i).$$

A GLM is fully defined by the distribution in the exponential family, the form of the linear predictor, and the response or link function. Some of the distributions in the exponential family include the normal, Poisson, Bernoulli, and binomial. Thus, GLMs provide a unified model for both continuous and discrete responses. In Chapter 2 we will show how the multinomial distribution can be embedded within the framework of a multivariate GLM.

In the linear mixed model (1.1) and the marginal linear mixed model (1.2) the expectations of the response $\mathbf{y}$ are the same. That is

$$E(\mathbf{y} \mid \mathbf{u}) = E(\mathbf{y}) = Z\boldsymbol{\beta}.$$

A simple example of this occurs in paired experiments where the mean difference across all subjects is the same as the difference between the two overall means. This desirable property unfortunately does not hold for non-normal response data. Thus in the analysis of longitudinal data for non-normal responses, a distinction is made between subject-specific (SS) models and population-averaged (PA) models (Zeger et al. 1988; Neuhaus et al. 1991; Agresti 1993b). In SS models the heterogeneity across subjects is modeled explicitly. Random or mixed effects models are examples of such models. Interpretation of the parameters in SS models refer to the influence of covariates upon individuals. In PA models the population-averaged response is modeled without explicitly accounting for the heterogeneity. The parameters in PA models are interpreted as the averaged population response to changes in the covariates. The relationship between the parameter estimates in the SS model and the PA model has been studied by Zeger et al. (1988) and Neuhaus et al. (1991) for the logistic and probit links and Ten Have et al. (1996) for the cumulative logit link. For instance in the logistic mixed model with a normal random intercept, the PA parameters $\boldsymbol{\beta}^*$ and the SS parameters $\boldsymbol{\beta}$, both of dimension p, satisfy $\mid \beta_k^* \mid \leq \mid \beta_k \mid$, $k = 1, \cdots, p$. In contrast the parameters in a log-linear model with a normal random intercept will have the same values for both the PA and SS approaches, expect for the intercept term (Diggle et al. 1994, p. 142).

To demonstrate the differences in inference between SS and PA modeling, consider the following example. Let y_{it} be the response at time t for subject i where $y_{it} = 1$ if the subject has high blood pressure and $y_{it} = 0$ otherwise. Let x_{it} be a corresponding dummy covariate denoting whether or not subject i was exercising at time t. From the PA model one would make inference about the rate of high blood pressure between exercisers and non-exercisers. The SS model would estimate the expected change in a subject's probability of having high blood pressure if they changed their exercise habits. Thus it is clear that the choice of PA versus SS modeling is often dependent on the study at hand and the desired inference. For population studies, such as those found in epidemiology, the PA approach typically provides the most informative inference. In contrast growth curve studies, where interest lies in a subject's response profile over time, lend themselves to the SS approach. In the next two sections we examine some of the modeling approaches that yield SS and PA inferences.

1.3.1 Subject-Specific Models

We first review GLMMs for Poisson, binomial, and Bernoulli data. As in the linear mixed model (1.1), GLMs are extended to GLMMs by incorporating random effects linearly within the linear predictor

$$\eta_{ij} = \mathbf{z}_{ij}'\boldsymbol{\beta} + \mathbf{w}_{ij}'\mathbf{u}_i. \tag{1.5}$$

Here the subscript j denotes the jth observation, $j = 1, \cdots, T_i$, for the ith subject, $\mathbf{z}_{ij}$ and $\mathbf{w}_{ij}$ are design vectors, and $\mathbf{u}_i$ is a vector of random effects. The GLMM is defined by first assuming that conditional on the random effects $\mathbf{u}_i$, y_{ij} satisfies the definition of a GLM with the linear predictor given in (1.5) and conditional mean $E(y_{ij} \mid \mathbf{u}_i) = \mu_{ij}$. The definition is completed by assuming the random effects $\mathbf{u}_i$ are independent with distribution $G(\mathbf{u}_i)$, and that the observations y_{ij} are conditionally independent within and between subjects.

If one is interested in making inference on both the fixed and random parameters, a full maximum likelihood approach to estimation should be used. In contrast one may be only interested in comparisons within subject. Thus a conditional likelihood approach could be used, where the random effects are treated as nuisance parameters and conditioned out of the likelihood (see, e.g., Conaway 1989).

Maximum likelihood estimation

In the full maximum likelihood approach the distribution of the random effects is incorporated into the likelihood. Though in theory the distribution, $G(\mathbf{u}_i)$, of the random effects may be any distribution, the common assumption is that $G(\mathbf{u}_i)$ is multivariate normal with $\mathbf{0}$ mean and covariance matrix Q. Estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, the elements of Q, entails maximization of the marginal likelihood of the y_{ij} obtained by integrating with respect to the random effects $\mathbf{u}_i$. Let $f(y_{ij} \mid \mathbf{u}_\mathrm{i})$ be the conditional distribution of y_{ij} given the random effect $\mathbf{u}_i$. For the GLMM (1.5), the marginal log-likelihood can be written

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int \cdots \int \prod_{j=1}^{T_i} f(y_{ij} \mid \mathbf{u}_\mathrm{i})\, G(\mathbf{u}_i)\, d\mathbf{u}_i. \tag{1.6}$$

Except for normal response data as in Section 1.2, the marginal likelihood will not have a closed form due to the intractable multivariate normal integrals.

The logistic-normal model (Pierce and Sands 1975) was one of the first random effects models for binary data in which the random term was incorporated on the same scale as the fixed effects. Both Pierce and Sands (1975) and Williams (1982) considered such a model that allowed for a random intercept (i.e. $w_{ij} = 1$ and $\mathbf{u}_i = u_i$ in model (1.5)). Williams (1982) proposed an iterative scheme for estimating the regression parameters and the variance of the random effect, which was based on a quasi-likelihood but only fit well when the variability was small. In what seems to be the first use of the term "generalized linear mixed model" in the literature, Gilmour et al. (1985) proposed an approximate method for fitting the general model (1.5) for the probit link. They proceeded by maximizing the likelihood with respect to the fixed effects, taking expectations over the random effects. The resulting estimating equations are analogs of the REML estimates of Harville (1977). Anderson and Aitkin (1985) suggested the use of numerical integration within the context of an EM algorithm for numerically approximating the marginal likelihood. They recommended the use of Gauss-Hermite quadrature, which will be discussed in detail in Chapter 3 along with an adaptive version of Gauss-Hermite quadrature (Liu and Pierce 1994; Pinheiro and Bates 1995). Hinde (1982) used a similar approach in the context of random Poisson regression models. The development of methods like this for calculating the exact maximum likelihood estimates by approximating the integrals using numerical or simulation techniques has since become an active area of research for GLMMs.

When the dimension of the random effects in model (1.5) is high, the use of numerical integration techniques such as Gauss-Hermite quadrature becomes computationally infeasible. A number of alternatives based on Monte Carlo (MC) methods have been proposed for handling the larger dimensional models. Zeger and Karim (1991) considered GLMMs in a Bayesian context and proposed an algorithm based on the Gibbs sampler. The Gibbs sampler is an MC method for generating observations from a complex joint posterior distribution, when sampling from the conditional distribution is easier. It involves a choice of prior distributions for the fixed parameters and the components of Q. Though computationally intensive, the algorithm can accommodate high dimensional integrals and can be easily modified to handle non-Gaussian random effect distributions. Karim and Zeger (1992) used this approach for analyzing the infamous salamander data set (McCullagh and Nelder 1989, p. 439-450) where the likelihood involved 40-dimensional intractable integrals. Though it has many attractive features, the use of a Bayesian paradigm with flat or diffuse priors to approximate the ML estimates may lead to posteriors that do not exist. This may not be detected when using the Gibbs sampler (Natarajan and McCulloch 1995; Hobert and Casella 1996), and could lead to incorrect estimates. McCulloch (1994) proposed a Monte Carlo EM (MCEM) algorithm to fit a probit-binomial model with normal random effects that used a Gibbs sampler to approximate the E-step. The Gibbs chain was based on the exact conditional distribution of $\mathbf{u}$ given $\mathbf{y}$. Chan and Kuk (1997) applied this approach to the salamander data set as well, allowing the random effects to be correlated.

McCulloch (1997) presented three algorithms for fitting GLMMs that rely on Monte Carlo techniques. The first algorithm, an MCEM algorithm, utilizes a Metropolis algorithm (Tanner 1993) to approximate the intractable integrals in the expectation step of the EM algorithm. In the Metropolis algorithm one chooses both a candidate distribution from which to sample new values, as well as an acceptance function that gives the probability of accepting the new values. McCulloch (1997) showed that by choosing $G(\mathbf{u}_i)$ as the candidate distribution, the acceptance function has a simple form involving only the joint conditional distribution $f(\mathbf{y} \mid \mathbf{u})$. The second algorithm again uses the Metropolis algorithm, but within the context of a Newton-Raphson algorithm. The algorithm iteratively solves a score equation for Q and score-type equation for $\boldsymbol{\beta}$. The scoring equation for $\boldsymbol{\beta}$ involves an intractable expectation where the Metropolis algorithm is applied. The final algorithm simulates the value of the likelihood directly as opposed to the previous algorithms that simulate the log-likelihood. Using a simulation study, McCulloch (1997) concluded that both the MCEM and the MC Newton-Raphson algorithms were feasible methods for calculating ML estimates in GLMMs. The simulated maximum likelihood (SML) approach performed poorly as a stand-alone algorithm. However, by running SML at the final estimates of the MCEM or MC Newton-Raphson algorithms, convergence issues can be addressed, slightly more precise estimates can be obtained, and an estimate of the maximized likelihood is available (McCulloch 1997).

A deficiency in the previous approaches based on the Gibbs and Metropolis algorithms is that the generated samples are dependent and thus MC error is difficult to assess. Without assessment of the MC error, one can not determine the appropriate number of simulations to use at each iteration of the algorithm. Booth and Hobert (1999) proposed an automated MCEM algorithm which utilized random sampling to construct the MC approximations at each E-step. With random sampling they were able to use standard central limit theory with Taylor series methods to assess the MC error at each iteration. The MC error was then used to determine the MC sample size, creating an automated algorithm. We will look at this algorithm in detail in Chapter 3.

Besides exact maximum likelihood analysis, there have been several proposed methods for carrying out approximate inference in GLMMs. Breslow and Clayton (1993) proposed a penalized quasi-likelihood (PQL) approach, Wolfinger and O'Connell (1993) utilized a pseudo-likelihood approach, while Engel and Keen (1994) used a combination of quasi-likelihood and REML. Though each motivated their methods using different approaches, they all yield equivalent estimates in certain cases, so we briefly describe the pseudo-likelihood approach. We first express model (1.5) in terms of the complete data

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = Z\boldsymbol{\beta} + W\mathbf{u}. \tag{1.7}$$

Wolfinger and O'Connell (1993) showed that the pseudo observation vector

$$\mathbf{v} = g(\hat{\boldsymbol{\mu}}) + g'(\hat{\boldsymbol{\mu}})\,(\mathbf{y} - \hat{\boldsymbol{\mu}}) \approx g(\mathbf{y}), \tag{1.8}$$

which is a Taylor series approximation to the linked response function $g(\mathbf{y})$, has distribution

$$\mathbf{v} \mid \boldsymbol{\beta}, \mathbf{u} \sim N\left[Z\boldsymbol{\beta} + W\mathbf{u}, g'(\hat{\boldsymbol{\mu}})\, R_{\hat{\boldsymbol{\mu}}}^{1/2}\, R\, R_{\hat{\boldsymbol{\mu}}}^{1/2}\, g'(\hat{\boldsymbol{\mu}})\right]. \tag{1.9}$$

Here $R_{\hat{\boldsymbol{\mu}}}$ is a known diagonal covariance matrix for the GLM under consideration, and R is an additional covariance matrix for modeling PA effects. In the PQL approach, R is the identity matrix. Considering $\boldsymbol{\beta}$ as unknown parameters and assuming $G(\mathbf{u}_i)$ is multivariate normal as before, then (1.9) takes the form of a weighted LMM with diagonal weight matrix

$$\hat{H} = R_{\hat{\boldsymbol{\mu}}}^{-1}\, [g'(\hat{\boldsymbol{\mu}})]^{-2}.$$

If we write $V = H^{-1/2}\, R\, H^{-1/2} + W\, Q\, W'$ then the solutions for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}_i$ take the same form as (1.3) and (1.4). REML or ML estimation can be used to estimate the elements of Q. There are a number of advantages to using the approximate methods for fitting GLMMs. By using approximations they avoid the intractable integrals that plague the exact maximum likelihood approaches. As a consequence the fitting of these models is relatively simple, and as seen above, existing methods for LMMs can be used. Unfortunately it has been shown that these methods perform poorly when the response is far from normal, as in binomial data with small sample sizes (Breslow and Clayton 1993; Breslow and Lin 1995; Lin and Breslow 1996).

The GLMMs discussed so far have assumed that the distribution of the random effects was normal and thus (1.6) involved intractable integrals. For certain models, however, the distribution of $G(\mathbf{u}_i)$ can be chosen such that (1.6) has a closed form. A number of such models (e.g., beta-binomial and Poisson-gamma) are found in the literature on overdispersion (see, e.g., Hinde and Demetrio 1998). Typically these models are not well suited for full random effects modeling as inclusion of between and within covariates is often difficult. Conaway (1990) proposed a random effects model for binary data that utilized the log-log link: $g(\mu_i) = \log(-\log(\mu_i))$. By assuming that the random effects distribution was log-gamma, the resulting marginal likelihood was shown to have a closed form. Conaway (1990) noted that the log-gamma distribution was flexible enough to resemble a variety of distributions including a normal distribution.

There has been some evidence to suggest that changes in the random effects distribution can lead to changes in the fixed effects parameter estimates (Heckman and Singer 1984; Davies 1987). Neuhaus et al. (1992) showed that model parameters were indeed inconsistent if the random effects distribution was misspecified, though the magnitude of the bias was typically small. There has been a moderate amount of recent work focused on nonparametric approaches to fitting random effects models (Davies 1987; Wood and Hinde 1987; Follmann and Lambert 1989; Aitkin 1996; Aitkin 1999). Such approaches assume that $G(\mathbf{u}_i)$ is a discrete distribution with unknown masses, mass points, and support size. Thus the integrals in (1.6) are replaced with a finite sum over an unknown support size. By maximizing (1.6) one obtains nonparametric maximum likelihood (NPML) estimates for the regression parameters as well as the discrete mixing distribution. In Chapter 4 we will consider an NPML approach for modeling nominal and ordinal response data.

The development of random effects models for nominal and ordinal data has lagged behind that for binomial data. Most of the proposed models have used either the cumulative logit or cumulative probit links for analyzing ordinal data. Harville and Mee (1984) proposed a cumulative probit random effects model and used an EM-type algorithm for obtaining estimates of fixed and random parameters. Using a Bayesian approach, they assumed a vague prior for the fixed parameters and applied a Taylor series approximation to evaluate the intractable normal integrals. Following the approach of Anderson and Aitkin (1985), Jansen (1990) used Gauss-Hermite quadrature to approximate the E-step in an EM algorithm. He considered the cumulative probit link and allowed for a single random intercept in a model for analyzing an agricultural experiment. Ezzet and Whitehead (1991) applied an ordinal random effects model to the analysis of a crossover experiment. A cumulative logit model was fit with a single random intercept that was assumed to be normally distributed. The Newton-Raphson method along with numerical integration were used to maximize the likelihood.

Two recent papers by Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996) presented general approaches for fitting GLMMs for ordinal response data, each using quite different estimation algorithms. Hedeker and Gibbons (1994) considered both cumulative probit and cumulative logit models and based their estimation methods on directly maximizing the marginal likelihood. To evaluate the intractable marginal likelihood, they approximated the normal integrals using multivariate Gauss-Hermite quadrature. Fisher's scoring method was used to obtain estimates of the parameters and the inverse of the expected information matrix was calculated at convergence to obtain standard errors. Tutz and Hennevogl (1996) motivated their model as a generalization of a multivariate GLM. Considering a cumulative logit model, they proposed three estimation procedures, each based on the EM algorithm. In two of the algorithms MC methods were used to approximate the integrals in the E-step, while in the third multivariate Gauss-Hermite quadrature was applied. Besides allowing for a random threshold models in which all threshold are shifted by the same random effect, Tutz and Hennevogl (1996) also proposed a model in which each threshold was allowed to vary according to its own random effect. They allowed the separate random effects to be either independent or correlated. We will consider this extended model in more detail in Chapter 3, as well as the EM algorithms proposed by Tutz and Hennevogl (1996). Based on a motivation similar to that in Tutz and Hennevogl (1996), Fahrmeir and Tutz (1994) proposed a nominal random effects regression model which utilized the baseline-category logit link. This model will be considered as well in Chapter 3.

As an extension of their previous methods for approximate inference in GLMMs (Engel and Keen 1994), Keen and Engel (1997) proposed an iteratively re-weighted REML estimation routine for mixed ordinal regression models. Their method was essentially the PQL approach of Breslow and Clayton (1993) for ordinal data using a cumulative logit link. We consider approximate inference methods for nominal and ordinal data in Chapter 3 based on generalizing the methods of Wolfinger and O'Connell (1993).

Two ordinal models that did not use the cumulative logit or probit links were proposed by Ten Have and Uttal (1994) and Ten Have (1996). Ten Have and Uttal (1994) proposed both PA and SS continuation-ratio logit models for analyzing multiple discrete time survival profiles of subjects in a psychological study. For a given set of multinomial probabilities $\pi_1, \cdots, \pi_R$ such that $\sum_{i=1}^{R} \pi_i = 1$, the continuation ratio logits are given as

$$L_j = \log\left(\frac{\pi_j}{\pi_{j+1} + \cdots \pi_R}\right), \quad j = 1, \cdots, R-1.$$

Ten Have and Uttal (1994) used a Bayesian approach assuming a non-informative prior on the regression parameters and a multivariate normal distribution for the random effects. They applied the Gibbs sampling approach of Zeger and Karim (1991) to estimate the model parameters. Ten Have (1996) extended the binary random effects model of Conaway (1990) to accommodate ordinal data. Here the cumulative complementary log-log link along with a log-gamma random effects distribution was used, which resulted in a closed form marginal likelihood.

Conditional maximum likelihood estimation

We now briefly describe the conditional maximum likelihood (CML) approach and mention a few applications. See, for example, Andersen (1980), Collett (1991) or Diggle et al. (1994) for further discussion. In CML estimation, one treats the random effects as nuisance parameters and conditions on their sufficient statistic. Estimation proceeds by maximizing the conditional likelihood. CML estimation is appropriate when one is interested in within subject comparisons. This approach is frequently used in matched case-control studies and in educational testing studies. An advantage of this approach over the ML approach is that one does not need to assume a distribution for the random effects. A consequence of conditioning, however, is that no information about the variability between subjects is obtained. Also, construction of sufficient statistics is limited to canonical link models, such as the logit model for binary data.

An example of the CML approach can be shown with item-response models. Such models arise in educational testing where a set of n subjects are administered a series of T questions (items) which have either a correct or incorrect answer. The Rasch model (Rasch 1961) assumes that the probability π_{1ij} of a correct answer for subject i and question j can be modeled by

$$\log\left(\frac{\pi_{1ij}}{1-\pi_{1ij}}\right) = \alpha_i - \beta_j. \tag{1.10}$$

Since there is a parameter α_i for each subject, the number of parameters increases as the sample size increases. In fact, ordinary ML estimators of β_j are inconsistent for model (1.10) (Andersen 1980, p. 244) . One could assume a distribution for the α_i and proceed by methods of the previous section to obtain marginal ML estimates of the β_j. Alternatively one can apply the CML approach by finding sufficient statistics for the α_i and then maximize the likelihood conditional on the sufficient statistics. As was originally shown by Tjur (1982), the CML estimates for model (1.10) correspond to standard ML estimates for certain log-linear models. It was later noted that the CML estimates of β_j were in fact the main effect ML estimates from the quasi-symmetry log-linear model (see, e.g., Fienberg 1986). Methods for fitting such models for nominal and ordinal responses have been discussed in Agresti (1993a, 1993b) and Agresti and Lang (1993).

1.3.2 Population-Averaged Models

For longitudinal or cluster data, models that study the averaged response over all subjects or clusters are called PA or marginal models. In these models, the relationship between the response and the explanatory variables is modeled separately from the association among repeated observations for an individual. Two of the landmark papers for PA modeling for GLMs are Zeger and Liang (1986) and Liang and Zeger (1986). To specify a marginal model one must specify the marginal mean, the marginal variance, and the covariance function. As an example we consider the marginal specification of a binary response model.

As in the definition of a GLM, we specify the marginal means as

$$\pi_{ij} = P(y_{ij} = 1 \mid \mathbf{z}_{ij}) = h_j(\mathbf{z}'_{ij}\boldsymbol{\beta}_j). \tag{1.11}$$

For the binary response model, the marginal variance depends on the marginal mean by

$$\operatorname{var}(y_{ij} \mid \mathbf{z}_{ij}) = \pi_{ij}(1 - \pi_{ij}). \tag{1.12}$$

In marginal models, one also must define a covariance function for modeling the covariance between y_{ij} and $y_{ij'}$:

$$\operatorname{cov}(y_{ij}, y_{ij'}) = c(\pi_{ij}, \pi_{ij'}, \boldsymbol{\alpha}). \tag{1.13}$$

The covariance function $c(\pi_{ij}, \pi_{ij'}, \boldsymbol{\alpha})$ depends on the marginal means and possibly additional association parameters $\boldsymbol{\alpha}$. A special feature of marginal models is that the parameters $\boldsymbol{\beta}$ can be consistently estimated even if $c(\pi_{ij}, \pi_{ij'}, \boldsymbol{\alpha})$ is misspecified (Zeger et al. 1988). Because of this, the $\operatorname{cov}(\mathbf{y}) = \Sigma(\boldsymbol{\beta}, \boldsymbol{\alpha})$, which denotes the combination of (1.12) and (1.13) for the complete data, is treated as a working covariance matrix.

A number of ways have been proposed for specifying the working covariance matrix. Liang and Zeger (1986) define $\Sigma(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in terms of a working correlation matrix R_α such that

$$\Sigma(\boldsymbol{\beta}, \boldsymbol{\alpha}) = R_{\boldsymbol{\pi}}^{1/2} \, R_{\boldsymbol{\alpha}} \, R_{\boldsymbol{\pi}}^{1/2},$$

where $R_{\boldsymbol{\pi}}$ is $\text{diag}[\pi_{ij}(1 - \pi_{ij})]$. If $R_{\boldsymbol{\alpha}} = I$ then all observations are treated as independent. One can parameterize $R_{\boldsymbol{\alpha}}$ to allow for, for example, autoregressive (AR) correlations, banded correlations, or entirely unstructured correlations. Lipsitz et al. (1991) proposed specifying the working covariance matrix in terms of odds ratios. One advantage of this approach is that the odds ratios are not constrained by the means π_{ij} as in the correlation specification.

Marginal model estimation is fundamentally different from SS model estimation since specification of (1.11), (1.12), and (1.13) does not, except for Gaussian data, fully define a likelihood. Estimation in marginal models is based on generalized estimating equations (GEEs) which are multivariate analogues of quasi-score functions (Wedderburn 1974). Briefly, in quasi-likelihood models the assumption of an exponential family is dropped from the GLM definition, and the model is defined only by assumptions on the first and second moments. Under appropriate conditions, parameters can be estimated consistently and asymptotic inference is still possible. The score functions for such models take the same form as those based on GLMs but are not likelihood equations because they lack a distributional assumption. Hence they are called quasi-likelihood or quasi-score functions. In matrix form for binary response data, the GEE for $\boldsymbol{\beta}$ is

$$s_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} Z_i' D_i \Sigma^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(\mathbf{y}_i - \boldsymbol{\pi}_i) = \mathbf{0} \qquad (1.14)$$

where Z_i is the design matrix for subject i and $D_i = \text{diag}\left[d\, h_j(\mathbf{z}_{ij}' \boldsymbol{\beta}_j)/d\eta_{ij}\right]$. For fixed $\boldsymbol{\alpha}$, estimates of $\boldsymbol{\beta}$ can be obtained from (1.14) by using Fisher's scoring.

Several ways have been suggested for estimating the unknown association parameters $\boldsymbol{\alpha}$. Liang and Zeger (1986) proposed using estimates based on Pearson residuals. Other approaches suggest defining an estimating equation for $\boldsymbol{\alpha}$ (see, e.g., Prentice 1988). The GEE approach has also been extended to the cases of repeated nominal and ordinal data (see, e.g., Lipsitz et al. 1994). Other marginal models that are in the same spirit as the GEE approach are the marginal quasi-likelihood (MQL) approach of Breslow and Clayton (1993), and the pseudo-likelihood approach of Wolfinger and O'Connell (1993) which contains a PA covariance matrix for modeling PA correlations.

We have stressed the differences in interpretation between the PA and SS approaches which arise from modeling the marginal mean in the former and the conditional mean, conditioned on the random effects, in the latter. The differences between PA and SS approaches have been blurred with a recent model proposed by Heagerty (1999). He proposed a marginalized latent variable model in which the marginal mean, not the conditional mean, conditioned on the random effects is modeled as a function of covariates. He presented their model from the context of a multi-level model which we outline here for the two level case. The model has two components, the first of which defines the regression structure for the marginal mean μ_{ij}:

$$g(\mu_{ij}) = \mathbf{z}'_{ij}\boldsymbol{\beta}^M, \tag{1.15}$$

while the second describes the dependence among measurements within a cluster:

$$g(\mu^u_{ij}) = \Delta(\mathbf{z}_{ij}) + u_i. \tag{1.16}$$

It is also assumed that the elements of the response vector $\mathbf{y}_i$ are conditionally independent given $\mathbf{u}$ and that the distribution of $\mathbf{u}$ is completely specified by the parameter $\boldsymbol{\alpha}$. The M in $\boldsymbol{\beta}^M$ is used to denote that the parameters are marginally defined. The parameter $\Delta(\mathbf{z}_{ij})$ in (1.16) is a function of both the linear predictor

$\eta_{ij} = \mathbf{z}_{ij}'\boldsymbol{\beta}^M$ and the random effects distribution $F_{\boldsymbol{\alpha}}(u_i)$. As an example, assume that $u_i \sim N[0, \sigma(\mathbf{z}_{ij})]$ and re-write $u_i = \sigma(\mathbf{z}_{ij})\, v_i$ where $v_i \sim N(0,1)$. The parameter $\Delta(\mathbf{z}_{ij})$ is then defined as the solution to the integral equation that links the marginal and conditional means:

$$\begin{aligned} \mu_{ij} &= E(\mu_{ij}^u) && (1.17) \\ h(\eta_{ij}) &= \int h\left[\Delta(\mathbf{z}_{ij}) + \sigma(\mathbf{z}_{ij})\, v_i\right] \phi(v_i)\, dv_i, && (1.18) \end{aligned}$$

where ϕ is the standard normal density function. Heagerty (1999) detailed an algorithm for fitting the marginalized latent variable model, which involves numerically evaluating the convolution equation (1.18) to solve for $\Delta(\mathbf{z}_{ij})$. Once obtained, existing maximum likelihood algorithms for GLMMs can be used to fit the marginalized models.

As noted before, $\boldsymbol{\beta}^M$ in (1.15) has a PA interpretation. From (1.18), $\sigma(\mathbf{z}_{ij})$ can be interpreted as a coefficient of a standardized omitted covariate v_i, with $\sigma(\mathbf{z}_{ij})$ contrasting subjects with equal $\Delta(\mathbf{z}_{ij})$ whose v_i differ by one unit (Heagerty 1999). It is also possible to calculate subject level effects based on the implied conditional linear predictor $\Delta(\mathbf{z}_{ij})$. Since the marginalized latent variable models are estimated by maximum likelihood and assume an underlying latent variable, they can be directly compared with conditionally defined models. The marginalized latent variable models provide the flexibility and interpretability of random effects models for introducing dependence while building regression structures for the marginal mean. Heagerty (1999) argued that marginal mean modeling allows for valid application with both time-dependent and time-independent covariates.

1.4 Outline

The models we will consider in Chapters 3 through 5 can be considered as multivariate generalized linear mixed models. Thus we will begin by defining multivariate

GLMs in Chapter 2 as well as introduce some notation that we will use throughout the dissertation. In Chapter 3 we will consider parametric approaches to modeling multinomial random effects models. We first define the multivariate generalized linear mixed model and then show how multinomial random effects models can be embedded within this framework. In this chapter we consider two approaches for modeling the multinomial random effects model, one based on numerical approximations of the integrals and the other based on Taylor series approximations. For the first approach we propose two algorithms, a quasi-Newton adaptive Gauss-Hermite algorithm and an automated Monte Carlo EM algorithm, while for the second we utilize a restricted pseudo-likelihood algorithm. A number of applications will be considered to illustrate the proposed models. We conclude Chapter 3 by examining the extended threshold model of Tutz and Hennevogl (1996).

As an alternative to the parametric approaches given in Chapter 3, we propose, in Chapter 4, a nonparametric approach for modeling multinomial random effects models. In particular, we consider models that allow for a random intercept and outline an EM algorithm for fitting such models. An important issue in nonparametric maximum likelihood methods is the identifiability of the model parameters. For the proposed multinomial models, we discuss this issue and provide a sufficient condition for ensuring identifiability. We illustrate the proposed models using one of the datasets analyzed in Chapter 3. We then conclude Chapter 4 with two simulation studies. In the first simulation study we compare the nonparametric maximum likelihood approach to parametric approaches in Chapter 3. In the second study we examine the performance of the Wald and likelihood-ratio tests for the nonparametric modeling approach as compared with the equivalent tests in the parametric approach.

In Chapter 5 we examine the use of the proposed methods in Chapters 3 and 4 for analyzing ordinal response data arising from multi-center clinical trials. In this type of data, two treatments are compared with respect to an ordinal response at multiple centers. If one assumes that the centers represent a random sample from some population of centers, one can utilize random effects models to account for heterogeneity among the centers. Typically, however, the number of centers is small and the assumption of normality for the random effects is questionable. We utilize simulations to examine the performance of a heterogeneous random effects model that includes a random center effect and a random center-by-treatment effect. We also propose an adaptive Gauss-Hermite quadrature approximated test for testing that a subset of the covariance matrix for the random effects is zero. We then conclude in Chapter 6 with a summary of the dissertation and proposals for possible areas of future research.

CHAPTER 2
MULTIVARIATE GENERALIZED LINEAR MODELS

2.1 Introduction

The models of the next three chapters will be motivated as extensions of multivariate generalized linear models. There are a number of advantages for motivating the models in this manner. First, approaching the models in this general framework allows for a single, unified notation for all models. Modifications for specific models involve relatively few changes, such as in the link function and design matrix. Second, the forms of the score functions, information matrices, and fitting algorithms are known for (multivariate) GLMs. Thus definitions of algorithms for the random effects models can be more easily derived using pieces from the fixed effects GLMs. Finally, extensions to other models for nominal and ordinal data not discussed here should be straightforward with the tools given in the next three chapters.

We begin in Section 2.2 by defining the multivariate GLM and by showing how the multinomial distribution can be embedded within the multivariate GLM framework. In Section 2.3 we discuss maximum likelihood estimation in multivariate GLMs for the special case of the multinomial distribution. In the final section we apply the general multivariate GLM framework to the specific models we will be discussing in the next three chapters. The notation introduced in the next three sections will be utilized throughout the remainder of the dissertation.

2.2 Definition

Proceeding as in Fahrmeir and Tutz (1994, Chap. 3), let $\mathbf{y}'_{ij} = (y_{ij1}, \cdots, y_{ijq})$ be a q-dimensional response vector with corresponding p-dimensional covariate vector $\mathbf{x}'_{ij}$ $= (x_{ij1}, \cdots, x_{ijp})$ where $E(\mathbf{y}_{ij} \mid \mathbf{x}_{ij}) = \boldsymbol{\mu}_{ij}$. In anticipation of the models in the next three chapters, we include both a subject i subscript, and an observation j within subject subscript. We also assume that subject i has T_i observations, $i = 1, \cdots, n$. For clarity we will use boldfaced lowercase letters (e.g., $\mathbf{y}$) for column vectors and denote matrices by capital letters.

As in the univariate GLM definition in Section 1.2, the multivariate GLM is defined by a distributional assumption and a structural assumption.

1. Distributional Assumption:

   The $\mathbf{y}_{ij}$ are assumed independent given the $\mathbf{x}_{ij}$ and have a distribution that belongs to a multivariate exponential family with the following form

   $$f(\mathbf{y}_{ij} \mid \boldsymbol{\theta}_{ij}, \phi, \boldsymbol{\omega}_{ij}) = \exp\left\{\frac{[\mathbf{y}'_{ij}\boldsymbol{\theta}_{ij} - b(\boldsymbol{\theta}_{ij})]}{\phi}\boldsymbol{\omega}_{ij} + c(\mathbf{y}_{ij}, \phi, \boldsymbol{\omega}_{ij})\right\}, \tag{2.1}$$

   where $\boldsymbol{\theta}_{ij}$ is the natural parameter, ϕ is an additional scale parameter, $b(\cdot)$ and $c(\cdot)$ are functions determined by the member of the exponential family, and $\boldsymbol{\omega}_{ij}$ is a vector of weights.

2. Structural Assumption:

   The linear predictor $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta}$ is related to the expectation $\boldsymbol{\mu}_{ij}$ by the vector-valued response function $\mathbf{h} = (h_1, \cdots, h_q)$ such that $\boldsymbol{\mu}_{ij} = \mathbf{h}(\boldsymbol{\eta}_{ij})$. Z_{ij} is a $(q \times p)$ design matrix and $\boldsymbol{\beta}' = (\beta_1, \cdots, \beta_p)$ is a vector of unknown parameters. Alternatively $\mathbf{g}(\boldsymbol{\mu}_{ij}) = \boldsymbol{\eta}_{ij}$ where $\mathbf{g}$, the link function, is the inverse of the response function $\mathbf{h}$.

We also define $\boldsymbol{v}(\cdot)$ to be the vector-valued function that relates the natural parameter $\boldsymbol{\theta}_{ij}$ directly to the linear predictor $\boldsymbol{\eta}_{ij}$ such that

$$\boldsymbol{\theta}_{ij} = \boldsymbol{v}(\boldsymbol{\eta}_{ij}). \tag{2.2}$$

The form of the design matrix, parameter vector, and the response and link functions will depend on the models being fit. In Section 2.4 we will define these items for the models that we will examine in Chapters 3 through 5.

Models for nominal and ordinal data are based on the multinomial distribution. Let $Y_{ij}^{(s)}$, $s = 1, \cdots, n_{ij}$, represent n_{ij} categorical response variables having possible values $1, \cdots, R = q + 1$. To express the multinomial distribution in the multivariate exponential form we first re-express $Y_{ij}^{(s)}$ as a dummy vector $\mathbf{y}_{ij}^{(s)\prime} = (y_{ij1}^{(s)}, \cdots, y_{ijq}^{(s)})$ where

$$y_{ijr}^{(s)} = \begin{cases} 1 & \text{if } Y_{ij}^{(s)} = r, \quad r = 1, \cdots, q. \\ 0 & \text{otherwise} \end{cases}$$

Thus for n_{ij} independent repetitions, $\mathbf{y}_{ij} = \sum_{s=1}^{n_{ij}} \mathbf{y}_{ij}^{(s)}$ is distributed multinomial with parameters n_{ij} and $\boldsymbol{\pi}_{ij}' = (\pi_{ij1}, \cdots, \pi_{ijq})$. Then, following Fahrmeir and Tutz (1994, p. 69), the distributional form of $\bar{\mathbf{y}}_{ij} = \mathbf{y}_{ij}/n_{ij}$ can be written

$$f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\theta}_{ij}, \phi, \boldsymbol{\omega}_{ij}) = \exp\left\{ \frac{[\bar{\mathbf{y}}_{ij}'\boldsymbol{\theta}_{ij} - b(\boldsymbol{\theta}_{ij})]}{\phi} \boldsymbol{\omega}_{ij} + c(\mathbf{y}_{ij}, \phi, \boldsymbol{\omega}_{ij}) \right\}, \tag{2.3}$$

where the natural parameter $\boldsymbol{\theta}_{ij}$ has components

$$\begin{aligned} \theta_{ijr} &= \log\left(\frac{\pi_{ijr}}{1 - \pi_{ij1} - \cdots - \pi_{ijq}} \right), \quad r = 1, \cdots, q, \\ b(\boldsymbol{\theta}_{ij}) &= -\log(1 - \pi_{ij1} - \cdots - \pi_{ijq}), \\ c(\mathbf{y}_{ij}, \phi, \boldsymbol{\omega}_{ij}) &= \log\left(\frac{n_{ij}!}{y_{ij1}! \cdots y_{ijq}!\,(n_{ij} - y_{ij1} - \cdots - y_{ijq})!} \right), \end{aligned}$$

and $\boldsymbol{\omega}_{ij} = n_{ij}$. In this framework the expectation $\boldsymbol{\mu}_{ij}$ is denoted by $\boldsymbol{\pi}_{ij}$.

2.3 Maximum Likelihood Estimation

We now outline maximum likelihood estimation for a multivariate GLM based on the multinomial form (2.3) (Fahrmeir and Tutz 1994, p. 98). In general, the MLE $\hat{\boldsymbol{\beta}}$ is calculated by finding the solution of the likelihood or score equations. This solution is only a local maxima in general, but corresponds to the global maximum as well when the log-likelihood is concave. The score equations are typically nonlinear and thus an iterative procedure for finding $\hat{\boldsymbol{\beta}}$ must be used. Common iterative procedures for fitting GLMs are Fisher scoring, iteratively re-weighted least squares, and Newton-Raphson, which we now describe.

We proceed by first calculating the score equations for $\boldsymbol{\beta}$, which require the first derivative of the log-likelihood. The log-likelihood for (2.3) depends on $\boldsymbol{\beta}$ only through the kernel

$$l_{ij}(\boldsymbol{\beta}) = \frac{\bar{\mathbf{y}}_{ij}'\boldsymbol{\theta}_{ij} - b(\boldsymbol{\theta}_{ij})}{\phi}\boldsymbol{\omega}_{ij}, \tag{2.4}$$

such that the log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{T_i} l_{ij}(\boldsymbol{\beta}). \tag{2.5}$$

In (2.4), $\boldsymbol{\theta}_{ij} = \boldsymbol{\theta}(\boldsymbol{\pi}_{ij})$ is a function of $\boldsymbol{\pi}_{ij}$, while $\boldsymbol{\pi}_{ij} = \boldsymbol{\pi}_{ij}(\boldsymbol{\beta}) = \mathbf{h}(Z_{ij}\boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$. Thus the derivative of (2.5) with respect to $\boldsymbol{\beta}$ requires the use of the chain rule for differentiation of vectors. Noting that $d\boldsymbol{\pi}_{ij}/d\,\boldsymbol{\beta} = Z_{ij}'\,D_{ij}$ where $D_{ij} = d\mathbf{h}/d\boldsymbol{\eta}$ evaluated at $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta}$, the score function takes the form

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{T_i} Z_{ij}'\,D_{ij}\,R_{\boldsymbol{\pi}_{ij}}^{-1}\,(\bar{\mathbf{y}}_{ij} - \boldsymbol{\pi}_{ij}). \tag{2.6}$$

In general $R_{\boldsymbol{\pi}_{ij}} = \mathrm{cov}(\mathbf{y}_{ij})$ is the covariance matrix for observation $\mathbf{y}_{ij}$ which depends on $\boldsymbol{\beta}$ through $\boldsymbol{\pi}_{ij}$. Note that (2.6) has the same form as the GEE for $\boldsymbol{\beta}$ (1.14), which was mentioned in Section 2.2.2. For the multivariate distribution, the covariance matrix for $\bar{\mathbf{y}}_{ij}$ has the form $R_{\boldsymbol{\pi}_{ij}} = \frac{1}{n_{ij}}(\mathrm{diag}(\boldsymbol{\pi}_{ij}) - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}_{ij}')$.

To obtain the expected Fisher information matrix, we take the expected value of $s(\boldsymbol{\beta})s(\boldsymbol{\beta})'$ yielding

$$\begin{aligned} F_E(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\sum_{j=1}^{T_i} Z_{ij}'\,D_{ij}\,R_{\boldsymbol{\pi}_{ij}}^{-1}\,D_{ij}'\,Z_{ij} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{T_i} Z_{ij}'H_{ij}\,Z_{ij}, \end{aligned} \tag{2.7}$$

where

$$\begin{aligned} H_{ij} &= D_{ij}\, R^{-1}_{\boldsymbol{\pi}_{ij}}\, D'_{ij} \\ &= \left\{ \frac{d\mathrm{g}(\boldsymbol{\pi}_{ij})}{d\boldsymbol{\pi}'}\, R_{\boldsymbol{\pi}_{ij}}\, \frac{d\mathrm{g}(\boldsymbol{\pi}_{ij})}{d\boldsymbol{\pi}} \right\}^{-1} \end{aligned}$$

is called the weight matrix.

Using (2.6) and (2.7), the Fisher scoring algorithm for estimation of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + F_E^{-1}(\hat{\boldsymbol{\beta}}^{(k)})\, s(\hat{\boldsymbol{\beta}}^{(k)}), \quad k = 0, 1, 2, \cdots . \tag{2.8}$$

This is in fact equivalent to iteratively re-weighted least squares. First note that we can re-write the score function (2.6) in terms of the weight matrix H_{ij} as

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{T_i} Z'_{ij}\, H_{ij}\, \frac{d\mathrm{g}(\boldsymbol{\pi}_{ij})}{d\boldsymbol{\pi}} \left[\bar{\mathbf{y}}_{ij} - \boldsymbol{\pi}_{ij}\right]. \tag{2.9}$$

If we then define the "pseudo" observation as

$$\check{\mathbf{y}}_{ij} = Z_{ij}\boldsymbol{\beta} + [D_{ij}^{-1}]'(\bar{\mathbf{y}}_{ij} - \boldsymbol{\pi}_{ij}),$$

we can re-write (2.8), using (2.9) and $\check{\mathbf{y}}' = (\check{\mathbf{y}}_{11}, \cdots, \check{\mathbf{y}}_{nT_n})$, as

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{(k+1)} &= F_E^{-1}(\hat{\boldsymbol{\beta}}^{(k)}) Z'\, H(\hat{\boldsymbol{\beta}}^{(k)})\, \check{\mathbf{y}}(\hat{\boldsymbol{\beta}}^{(k)}) \\ &= \left[Z'\, H(\hat{\boldsymbol{\beta}}^{(k)})\, Z\right]^{-1} Z'\, H(\hat{\boldsymbol{\beta}}^{(k)})\, \check{\mathbf{y}}, \end{aligned} \tag{2.10}$$

which is of the form of a weighted least-squares estimate. By iteratively calculating (2.10), one can obtain the MLE for $\boldsymbol{\beta}$. Here Z and H denote the design and weight matrices, respectively, for the entire data, where

$$Z = \begin{bmatrix} Z_{11} \\ Z_{12} \\ \vdots \\ Z_{nT_n} \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} H_{11} & & & 0 \\ & H_{12} & & \\ & & \ddots & \\ 0 & & & H_{nT_n} \end{bmatrix}.$$

For the remainder of this dissertation, when defining block diagonal matrices we will use the notation $H = \text{diag}(H_{ij})$ and for stacked matrices or vectors we will use $Z = [Z_{ij}]$. Thus we could define the complete pseudo observation vector as $\check{\mathbf{y}} = [\check{\mathbf{y}}_{ij}]$.

The Newton-Raphson algorithm is of the same form as (2.8), but with the expected information matrix F_E replaced by the observed information matrix F_{obs}. The observed information matrix is defined as the negative of the second derivative of the log-likelihood with respect to $\boldsymbol{\beta}$. For canonical link functions, where $\mathbf{g}(\boldsymbol{\mu}) = \boldsymbol{\theta}$, the observed and expected information matrices coincide. In general, the contribution of the jth observation from the ith subject to the observed information matrix is, suppressing dependence on $\boldsymbol{\beta}$,

$$F_{O,ij} = F_{E,ij} - O_{ij}, \tag{2.11}$$

where

$$O_{ij} = O_{ij}(\boldsymbol{\beta}) = \sum_{r=1}^{q} Z'_{ij} U_{ijr}(\boldsymbol{\beta}) Z_{ij}(\bar{y}_{ijr} - \pi_{ijr}). \tag{2.12}$$

The matrix U_{ijr} is the matrix of second derivatives

$$\frac{d^2 v_r(\boldsymbol{\eta}_{ij})}{d\boldsymbol{\eta} d\boldsymbol{\eta}'}$$

where $\boldsymbol{v}(\cdot)$ was defined in (2.2). For models based on the logit link, U_{ijr} has the form

$$\begin{aligned}\frac{d^2 v_r(\boldsymbol{\eta}_{ij})}{d\boldsymbol{\eta} d\boldsymbol{\eta}'} = \frac{1}{h_r}\frac{d^2 h_r}{d\boldsymbol{\eta} d\boldsymbol{\eta}'} - \frac{1}{h_r^2}\frac{dh_r}{d\boldsymbol{\eta}}\frac{dh_r}{d\boldsymbol{\eta}'} + \frac{1}{1-\sum_{l=1}^q h_l}\frac{d^2[\sum_{l=1}^q h_l]}{d\boldsymbol{\eta} d\boldsymbol{\eta}'} \\ + \frac{1}{(1-\sum_{l=1}^q h_l)^2}\frac{d[\sum_{l=1}^q h_l]}{d\boldsymbol{\eta}}\frac{d[\sum_{l=1}^q h_l]}{d\boldsymbol{\eta}'},\end{aligned} \tag{2.13}$$

where h_r is the rth component of $\mathbf{h}(\boldsymbol{\eta}_{ij})$. Note that $\dfrac{dh_r}{d\boldsymbol{\eta}}$ is the rth column of D_{ij}. To evaluate (2.13), one also needs the q by q matrix of second derivatives $\frac{d^2 h_r}{d\boldsymbol{\eta} d\boldsymbol{\eta}'}$.

2.4 Applications

Though we will present the random effects approaches in the next three chapters in a general form, we will examine a number of specific models. For nominal data we will consider the baseline-category logit model, while for ordinal data we will examine the cumulative logit, adjacent-category logit, and the continuation-ratio logit models.

2.4.1 Baseline-Category Logit Model

Categorical variables that do not have a natural ordering for their levels are called nominal variables. A statistical model that is appropriate for assessing the influence of explanatory variables on a nominal response is one that uses baseline-category logits (Agresti 1990, p. 307). We refer to such a model as the baseline-category logit model, though it could also be called the multinomial logit model or the polychotomous logistic regression model.

A common application of baseline-category logit models is in discrete-choice modeling (Maddala 1983). Such models often appear in the econometric literature. In discrete-choice models, subjects are presented with a set of R possible choices. Explanatory variables in discrete-choice models can be classified as either a characteristic of the chooser (e.g., gender, race, income) or a characteristic of the choice (e.g., price of item, color of item). We will see below how these two types of covariates influence the form of the design matrix.

For the baseline-category logit model, the response function $\mathbf{h}(\boldsymbol{\eta}_{ij})$ has components

$$h_r(\boldsymbol{\eta}_{ij}) = \frac{\exp(\eta_{ijr})}{1 + \sum\limits_{l=1}^{q} \exp(\eta_{ijl})}, \quad r = 1, \cdots, q. \tag{2.14}$$

Alternatively, the link function $\mathbf{g}(\boldsymbol{\pi}_{ij})$ has components

$$g_r(\boldsymbol{\pi}_{ij}) = \log \frac{\pi_{ijr}}{1 - \sum\limits_{l=1}^{q} \pi_{ijl}}. \tag{2.15}$$

Thus the q logits in the baseline-category logit model are formed by pairing each response category with a baseline response category, which we take to be the last one (category $R = q + 1$).

In the baseline-category logit model there is a separate parameter vector $\boldsymbol{\beta}_r$ for each of the q logits. Let $\mathbf{x}_{ij}$ be the covariates related to the chooser and let $\mathbf{v}_r$ be the covariates related to the choice. The subscript r refers to the rth choice and implies that covariates for the choice are the same across all subjects and observations. Since logits are formed by pairing responses to a baseline category, the design matrix Z_{ij} including both types of covariates has the form

$$Z_{ij} = \begin{bmatrix} 1 & \mathbf{x}'_{ij} & & & & & \mathbf{v}'_1 - \mathbf{v}'_R \\ & & 1 & \mathbf{x}'_{ij} & & & \mathbf{v}'_2 - \mathbf{v}'_R \\ & & & & \ddots & & \vdots \\ & & & & 1 & \mathbf{x}'_{ij} & \mathbf{v}'_q - \mathbf{v}'_R \end{bmatrix}. \tag{2.16}$$

For design matrix (2.16), the parameter vector is

$$\boldsymbol{\beta}' = (\boldsymbol{\beta}'_1, \cdots, \boldsymbol{\beta}'_q, \boldsymbol{\gamma}')$$

where $\boldsymbol{\beta}'_r = (\alpha_r, \tilde{\boldsymbol{\beta}}_r')$ contains a threshold parameter for the rth logit and a covariate parameter vector corresponding to the chooser covariates for the rth logit, $r = 1, \cdots, q$. The parameters $\tilde{\boldsymbol{\beta}}_r$ have log odds interpretations with respect to the baseline-category. The final parameter vector $\boldsymbol{\gamma}$ corresponds to the differences in the covariate vectors between the paired response categories. The parameter $\boldsymbol{\gamma}$ measures the influence of the characteristics of the choices and its interpretation is the same across all logits.

The derivative matrix $D_{ij} = d\mathbf{h}/d\boldsymbol{\eta}$ evaluated at $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta}$ is required for carying out maximum likelihood estimation. The (u, v)th element in the q by q matrix D_{ij} corresponds to the derivative $\dfrac{dh_v}{d\eta_{iju}}$. For the baseline-category logit model, D_{ij} takes on the familiar form

$$D_{ij} = \text{diag}(\boldsymbol{\pi}_{ij}) - \boldsymbol{\pi}_{ij}\boldsymbol{\pi}'_{ij}.$$

Since the link function is the same as the natural parameter for the baseline-category logit model, O_{ij} in (2.12) is zero.

2.4.2 Adjacent-Category Logit Model

We now consider models for ordinal data, the first being the adjacent-category logit model. When analyzing such response variables, a model should be chosen that will account for the ordering. The models that we will examine incorporate the ordering directly into the link functions. The adjacent-category logit model is a special case of a model originally considered by Andersen (1973) and is described in detail in Agresti (1990). As the name implies, adjacent categories of the ordinal response are used to form logits. When all covariates are categorical, the adjacent-category logit model corresponds to a log-linear model with scores assigned to the ordinal response. Unlike the baseline-category logit model, a common association parameter $\boldsymbol{\beta}$ is usually assumed to hold for all adjacent-category logits. This is how the model treats the response as ordinal.

The form of the response and link functions for the adjacent-category logit model is very similar to (2.14) and (2.15) for the baseline-category logit model. If we let $\eta^*_{ijr} = (r - R)\,\eta_{ijr}, r = 1, \cdots, q$, then the components of the response and link functions for the adjacent-category logit model have the same form as those for the baseline-category logit model. That is

$$h_r(\boldsymbol{\eta}^*_{ij}) = \frac{\exp(\eta^*_{ijr})}{1 + \sum\limits_{l=1}^{q} \exp(\eta^*_{ijl})}, \quad r = 1, \cdots, q, \tag{2.17}$$

and

$$g_r(\boldsymbol{\pi}^*_{ij}) = \log\left(\frac{\pi^*_{ijr}}{1 - \sum\limits_{l=1}^{q} \pi^*_{ijl}}\right), \tag{2.18}$$

where π^*_{ijr} is a function of the altered design matrix.

Since the adjacent-category logit model has a single association parameter, the types of covariates $\mathbf{x}_{ij}$ for each pair of logits must be the same. Therefore the design matrix is given by

$$Z_{ij} = \begin{bmatrix} 1 & & & & \mathbf{x}'_{ij} \\ & 1 & & & \mathbf{x}'_{ij} \\ & & \ddots & & \mathbf{x}'_{ij} \\ & & & 1 & \mathbf{x}'_{ij} \end{bmatrix}. \tag{2.19}$$

The parameter vector $\boldsymbol{\beta}' = (\alpha_1, \cdots, \alpha_q, \boldsymbol{\beta}^{*'})$ includes q threshold parameters and the parameters associated with the covariate vector. These parameters have log odds interpretations that hold across all adjacent pairs of responses. Calculated from the adjusted design matrix, the derivative matrix $D_{ij} = \text{diag}(\boldsymbol{\pi}^*_{ij}) - \boldsymbol{\pi}^*_{ij}\boldsymbol{\pi}^{*'}_{ij}$ and $O_{ij} = 0$.

2.4.3 Cumulative Logit Model

One of the most popular ordinal response models is the cumulative logit model (McCullagh 1980; Agresti 1990, sec. 9.4). The logits in the cumulative logit model are functions of cumulative probabilities, where the rth logit is a logit for a binary response in which categories 1 to r are paired with categories $r + 1$ to R. The association parameter $\boldsymbol{\beta}$ is usually assumed to be the same across all logits and it is interpreted as a cumulative log odds. The cumulative log odds ratio for covariates $\mathbf{x}_1$ and $\mathbf{x}_2$ then is proportional to the difference between the covariates, which holds across all logits. Because of this property, the model is often referred to as the proportional odds model. Since this model uses groupings of categories rather than individual categories, cumulative logit models are not equivalent to log-linear models.

The components of the response function for the cumulative logit model take the form

$$\begin{aligned} \pi_{ij1} &= h_1(\boldsymbol{\eta}_{ij}) = \frac{1}{1+\exp(-\eta_{ij1})} \\ \pi_{ijr} &= h_r(\boldsymbol{\eta}_{ij}) = \frac{1}{1+\exp(-\eta_{ijr})} - \frac{1}{1+\exp(-\eta_{ij,r-1})}, \quad r = 2, \cdots, q. \end{aligned} \tag{2.20}$$

and the components of the link function satisfy

$$g_r(\boldsymbol{\eta}_{ij}) \;=\; \log\left(\frac{\sum\limits_{l=1}^{r} \pi_{ijl}}{1 - \sum\limits_{l=1}^{r} \pi_{ijl}} \right). \quad r = 2, \cdots, q. \tag{2.21}$$

Since the parameter vector $\boldsymbol{\beta}$ is constant across all logits, the design matrix takes the same form as that for the adjacent-category logit model (2.19):

$$Z_{ij} = \begin{bmatrix} 1 & & & \mathbf{x}_{ij}^{'} \\ & 1 & & \mathbf{x}_{ij}^{'} \\ & & \ddots & \vdots \\ & & 1 & \mathbf{x}_{ij}^{'} \end{bmatrix}. \tag{2.22}$$

The parameter vector $\boldsymbol{\beta}' = (\alpha_1, \cdots, \alpha_q, \boldsymbol{\beta}^{*'})$ includes q threshold parameters and the parameters associated with the covariate vector. The thresholds are strictly stochastically ordered such that $\alpha_1 < \cdots < \alpha_q$. The cumulative logit model can be motivated from a latent variable approach (see Section 3.2), in which the true underlying response is measured only in relation to where the thresholds lie. For this reason, these models are also referred to as threshold models.

Let $\Pi_{ijr} = \pi_{ij1} + \cdots + \pi_{ijr}$ and let $\Psi_{ijr} = \Pi_{ijr}(1 - \Pi_{ijr}),\ r = 1, \cdots, q$. Then the derivative matrix D_{ij} for the cumulative logit model takes the form

$$D_{ij} = \begin{bmatrix} \Psi_{ij1} & -\Psi_{ij1} & & & & 0 \\ & \Psi_{ij2} & -\Psi_{ij2} & & & \\ & & \ddots & \ddots & & \\ & & & \Psi_{ij,q-1} & -\Psi_{ij,q-1} & \\ 0 & & & & & \Psi_{ijq} \end{bmatrix} . \tag{2.23}$$

For the cumulative logit model, $\mathbf{g}(\boldsymbol{\pi}_{ij}) \neq \boldsymbol{\theta}_{ij}$ and so the matrix O_{ij} is not zero. To calculate the observed information matrix, the second derivative matrix $\dfrac{d^2 h_r}{d\boldsymbol{\eta}_{ij} d\boldsymbol{\eta}_{ij}'}$ is needed for evaluating (2.12). For Π_{ijr} and Ψ_{ijr} defined as above, $\dfrac{d^2 h_r}{d\boldsymbol{\eta}_{ij} d\boldsymbol{\eta}_{ij}'}$ has (u, v)th element

$$\begin{aligned} \frac{d^2 h_r}{d\eta_{iju} d\eta_{ijv}} &= \Psi_{ijr}(1 - 2\Pi_{ijr}) && \text{if } u = v = r\ , \\ &= -\Psi_{ijr}(1 - 2\Pi_{ij,r-1}) && \text{if } u = v = r - 1, \\ &= 0 && \text{otherwise,} \end{aligned}$$

where Γ_{ij0} is defined to be zero.

2.4.4 Continuation-Ratio Logit Model

The final model for ordinal data is the continuation-ratio logit (CRL) model (Cox 1972, Agresti 1990, p. 319). Given a set of multinomial response probabilities, the continuation ratio is defined as the ratio of the rth multinomial probability over the sum of the remaining $r+1$ to R probabilities. Applications of the CRL model include, for example, modeling discrete time data (see, e.g., Ten Have and Uttal 1994) and modeling data in which the ordering of the response is due to a sequential mechanism

(Fahrmeir and Tutz 1994, sec. 3.3.4). An example of response data in which the ordered categories are defined sequentially is found in McCullagh (1980). The data set consisted of information on tonsil sizes of children. Each child was classified according to their relative tonsil size and if they had Streptococcus pyogenesis. The tonsil size was classified as "Present but not enlarged", "Enlarged", or "Greatly enlarged". Children are assumed to start in a normal state (category 1). If the tonsils start to grow abnormally, they could become enlarged (category 2). If they keep growing, they could move to the final group and be very enlarged. However, to get to the third category, the tonsil must move through the first two categories. Therefore the ordinal responses are sequential in nature. The CRL model can be motivated from underlying latent variables based on this sequential mechanism which will be outlined in Section 3.2.2.

For the CRL model, the components of the response function take the form

$$h_r(\boldsymbol{\eta}_{ij}) = \frac{1}{1+\exp(-\eta_{ijr})}\prod_{l=1}^{r-1}\left[1-\frac{1}{1+\exp(-\eta_{ijl})}\right], \quad r = 1, \cdots, q, \tag{2.24}$$

where $\prod_{r=1}^{0}\{\cdot\} = 1$, and the components of the link function are

$$g_r(\boldsymbol{\eta}_{ij}) = \log\frac{\pi_{ijr}}{\pi_{ij,r+1}+\cdots\pi_{ijq}}, \quad r = 1, \cdots, q. \tag{2.25}$$

If the parameters in each of the logits for the CRL models are distinct, then fitting the models separately for each logit with yield the same results as fitting all logits simultaneously (Agresti 1990, p. 319). For ordinal models, however, one can assume that the association parameter $\boldsymbol{\beta}$ is the same across all logits and has the form $\boldsymbol{\beta}' = (\alpha_1, \cdots, \alpha_q, \boldsymbol{\beta}^{*'})$. In contrast to the cumulative logit model, the thresholds α_r for the CRL model are not ordered. The design matrix for the model with common effect parameter has the familiar form

$$Z_{ij} = \begin{bmatrix} 1 & & & \mathbf{x}_{ij}' \\ & 1 & & \mathbf{x}_{ij}' \\ & & \ddots & \vdots \\ & & 1 & \mathbf{x}_{ij}' \end{bmatrix}. \tag{2.26}$$

The matrix of derivatives D_{ij} is somewhat complicated due to the form of the response function (2.24). Let $\Gamma_{ijr} = \dfrac{1}{1+\exp(-\eta_{ijr})}$. The (u, v)th element, d_{uv}, of D_{ij} has the form

$$\begin{aligned} d_{uv} &= 0 && \text{if } u > v, \\ &= \Gamma_{iju} \prod_{l=1}^{u} (1 - \Gamma_{ijl}) && \text{if } u = v, \\ &= -\Gamma_{iju}\Gamma_{ijv} \prod_{l=1}^{v-1} (1 - \Gamma_{ijl}) && \text{if } u < v, \end{aligned} \tag{2.27}$$

where $\prod_{r=1}^{0}\{\cdot\} = 1$. In addition, to calculate the observed information matrix, the (u, v)th element of the matrix of second derivatives $\dfrac{d^2 h_r}{d\eta_{ij} d\eta_{ij}'}$ is given by

$$\begin{aligned} \frac{d^2 h_r}{d\eta_{iju} d\eta_{ijv}} &= 0 && \text{if } r < u \text{ or } r < v, \\ &= \Gamma_{ijr}(1 - 2\Gamma_{ijr}) \prod_{l=1}^{r} (1 - \Gamma_{ijl}) && \text{if } u = v = r, \\ &= -\Gamma_{ijr}\Gamma_{iju}(1 - 2\Gamma_{iju}) \prod_{l=1}^{u} (1 - \Gamma_{ijl}) && \text{if } u = v \neq r, \\ &= -\Gamma_{ijr}\Gamma_{ijv} \prod_{l=1}^{r} (1 - \Gamma_{ijl}) && \text{if } u = r,\ v \neq r, \\ &= \Gamma_{ijr}\Gamma_{iju}\Gamma_{ijv} \prod_{l=1}^{v-1} (1 - \Gamma_{ijl}) && \text{if } u \neq v \neq r, \end{aligned}$$

where $\prod_{r=1}^{0}\{\cdot\} = 1$.

CHAPTER 3
MULTIVARIATE GENERALIZED LINEAR MIXED MODELS FOR NOMINAL AND ORDINAL RESPONSE DATA

3.1 Introduction

As stated in Chapter 1, there has been relatively little research in the area of random effects models for nominal and ordinal response data. The majority of this work has been focused on models for ordinal responses with cumulative logit or probit links that have allowed only simple random effects structures (Jansen 1990; Ezzet and Whitehead 1991), or have been based on Taylor series approximations (Harville and Mee 1984). Special models for correlated discrete failure time data with ordinal responses have also been considered by Ten Have and Uttal (1994) and Ten Have (1996) which utilized the continuation-ratio and complementary log-log links, respectively. In the former, estimation was carried out by way of the Gibbs sampling routine by first assuming a noninformative prior distribution for the regression parameters. Only recently has a general approach for modeling clustered ordinal response data been presented (Hedeker and Gibbons 1994; Tutz and Hennevogl 1996).

Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996) proposed similar models for repeated ordinal responses, though they considered quite different estimation routines. Hedeker and Gibbons (1994) considered a general random effects model for ordinal data using either the cumulative logit or probit links. They directly maximized the marginal likelihood obtained by approximating the normal integrals by Gauss-Hermite quadrature. Tutz and Hennevogl (1996) also proposed an ordinal regression model that allowed for a general random effects structure. They motivated their model as a multivariate generalized linear mixed model and considered Gauss-Hermite quadrature and Monte Carlo EM algorithms to maximize the log-likelihood.

Tutz and Hennevogl (1996) also proposed an ordinal model in which each threshold was assumed to be random. They relaxed the usual assumption in which all thresholds for a given subject are shifted randomly by the same amount, and allowed each threshold to vary according to its own distribution. Estimation in this general model is more difficult since the order restriction on the thresholds (see Section 2.4.3) may be violated if the variabilities of the thresholds are large or the thresholds are not well separated.

Random effects models for nominal response data have received even less attention in the statistical literature. Fahrmeir and Tutz (1994, p. 231) outlined a baseline-category logit model which allowed for random thresholds. Hedeker (2000) proposed a similar baseline-category model and provided a Fortran program that approximated the normal integrals using Gauss-Hermite quadrature. A notable deficiency in his proposed model, however, was the inability to estimate correlations between random effects in different thresholds. For example, Hedeker (2000) allowed each threshold for a given subject to vary according to its own distribution, but assumed that the threshold random effects were perfectly correlated. One would expect that thresholds from the same subject would be correlated, but assuming that the correlation is always 1.0 is an overly strong assumption.

Some special cases of the baseline-category logit random effects model have been examined in the psychometric literature. In psychometric research a popular model for analyzing item response data is the Rasch model (Rasch 1961). In such models it is assumed that the items (questions) being measured describe some underlying latent trait, or traits, of a set of cases (subjects). Questions on standardized tests, for example, are used to assess the underlying verbal or mathematical ability of students. When the items being measured have nominal responses, the baseline-category logit model can be used to estimate the item parameters. Adams and Wilson (1996) considered such a model, but allowed the thresholds to be shifted for each subject.

This model was then extended by Adams et al. (1997) to allow the thresholds to vary individually for each subject. For both models, an EM algorithm was used to maximize the marginal log-likelihood. Two approaches were used to obtain the marginal log-likelihood. In the first approach, the random effects were assumed to be multivariate normal, and a grid of points was chosen at which the multivariate normal density was approximated. The points and the approximate weights were then used to approximate the integrals. In the second approach, the random effects were assumed to follow a discrete step distribution defined on a prespecified set of nodes. The density values at the nodes of the discrete step distribution were estimated within the EM algorithm.

In this chapter we propose general random effects models for nominal and ordinal response data. We motivate these models as extensions of the multivariate generalized linear model considered in Section 2.1. The resulting multivariate generalized linear mixed model provides a unified framework from which various models for nominal and ordinal data can be motivated. In particular, we consider four multinomial logit models that have link functions based on the logit link. For nominal data we consider the baseline-category logit model and for ordinal data we consider the continuation-ratio logit model, the adjacent-category logit model, and the cumulative logit model. For the baseline-category logit model, we allow for a general random effects structure which includes correlated random effects between thresholds, in contrast to Hedeker (2000). This approach is more general then that of Adams and Wilson (1996) and Adams et al. (1997), in that it includes their models as special cases. We also generalize the work of Ten Have and Uttal (1994) on the continuation-ratio logit model by considering a general regression model for ordinal responses. The proposed model for the adjacent-category logit link, to our knowledge, has not been considered previously. Our random cumulative logit model is similar to that of Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996), but we employ a different estimation routine.

In particular, for all models we utilize adaptive multivariate Gauss-Hermite quadrature to numerically approximating the intractable multivariate normal integrals, and then proceed by directly maximizing the marginal log-likelihood. This approach has not been utilized before for multinomial response models. We also apply the Monte Carlo EM algorithm of Booth and Hobert (1999) as an alternative estimation routine for high dimensional random effects models.

In addition to the proposed maximum likelihood methods, we also generalize the work of Wolfinger and O'Connell (1993) to allow for approximate inference in mixed nominal and ordinal regression models. Keen and Engel (1997) proposed an iteratively re-weighted REML estimation routine for ordinal response data which used minimum norm quadratic estimation (MINQUE) (Rao 1973) to obtain estimates of the variance components. An advantage of the MINQUE estimation method is that it is noniterative, providing method of moment type estimators for the variance components. A disadvantage, however, is that MINQUE estimates can be negative. Swallow and Monahan (1984) recommended the use of ML or REML estimates over MINQUE based on results of a series of simulation studies. Our extension of the methods of Wolfinger and O'Connell (1993) is based on a pseudo-likelihood approach that utilizes REML estimation for the variance components. We again motivate the model in terms of a multivariate generalized linear mixed model, which allows for simple application to the links discussed above.

The remainder of the chapter is structured as follows: We begin in Section 3.2 by defining the multivariate generalized linear mixed model. Within that section we also consider the motivation of the nominal and ordinal mixed models as extensions of linear mixed models. In Section 3.3 we discuss the estimation methods for obtaining maximum likelihood estimates of the regression parameters and variance components. We provide details for obtaining estimates of the standard errors upon convergence of the algorithms and for carrying out inferences in Section 3.4. We then present in

Section 3.5 an approximate maximum likelihood method for fitting models for nominal and ordinal response data. In Section 3.6 we apply the models of this chapter to a number of datasets. We conclude in Section 3.7 by considering the extended random threshold model of Tutz and Hennevogl (1996).

3.2 Multivariate Generalized Linear Mixed Models

Multivariate generalized linear mixed models (MGLMMs) are extended multivariate generalized linear models that incorporate random effects linearly along with the fixed effects in the linear predictor. In this section we define MGLMMs and show that multinomial random effects models are special cases of MGLMMs. We then show that the multinomial random effects models under consideration can be motivated from an underlying latent variable that follows a linear mixed model.

3.2.1 Definition

As in Section 2.1, let $\mathbf{y}'_{ij} = (y_{ij1}, ..., y_{ijq})$ be a q-dimensional response vector with corresponding p-dimensional covariate vector $\mathbf{x}'_{ij} = (x_{ij1}, ..., x_{ijp})$, $j = 1, \cdots, T_i$, $i = 1, \cdots, n$, and denote the fixed effects parameter vector by $\boldsymbol{\beta}$. Also, for the ith subject let $\mathbf{u}'_i = (u_{i1}, \cdots, u_{im})$ be an m-dimensional vector of subject-specific random effects. The multivariate generalized linear mixed model is defined by the following two-stage model.

1. Stage 1:

   Assume that given the random effects $\mathbf{u}_i$, the distribution, $f(\mathbf{y}_{ij} \mid \mathbf{u}_i; \boldsymbol{\beta})$, of $\mathbf{y}_{ij}$ is a member of the multivariate exponential family with conditional mean

   $$\boldsymbol{\mu}_{ij} = E(\mathbf{y}_{ij} \mid \mathbf{u}_{ij}) = \mathbf{h}(\boldsymbol{\eta}_{ij}) \text{ and } \boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W_{ij}\mathbf{u}_i, \tag{3.1}$$

   where the response function $\mathbf{h}$ and the design matrix Z_{ij} are defined as in Chapter 2, and W_{ij} denotes the design matrix for the covariates that are assumed to vary across subjects.

2. Stage 2:

   Assume that the subject-specific random effects $\mathbf{u}_i$ are independent and follow a distribution G with mean $\mathbf{0}$ and positive definite variance-covariance matrix Σ.

In addition, it is assumed that observations within a subject are conditionally (on $\mathbf{u}_i$) independent, and observations between subjects are conditionally and unconditionally independent. Thus the conditional density of the complete response vector $\mathbf{y}$ and random effects vector $\mathbf{u}$ can be written

$$f(\mathbf{y} \mid \boldsymbol{\beta}; \mathbf{u}) = \prod_{i=1}^{n} \prod_{j=1}^{T_i} f(\mathbf{y}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i).$$

Note that Stage 1 requires that conditional on $\mathbf{u}_i$, $\mathbf{y}_{ij}$ follows a multivariate generalized linear model as defined in Section 2.1.

The usual assumption that is made concerning G, and the one that we make in this chapter, is that G is the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. In general, however, any distribution can be chosen for G. For example, in the next chapter we consider G to be a discrete distribution with unknown support size, masses, and mass points. The assumption of normality is popular as it allows for a variety of covariance structures for the random effects. It does, however, create difficulties for estimation of the fixed and random effects since this entails maximizing the marginal likelihood of the data. Obtaining the marginal likelihood is hampered by the intractable integrals of the normal distribution. In Section 3.3 we present an estimation routine that uses adaptive Gauss-Hermite quadrature to obtain the marginal likelihood.

As was shown in Section 2.1, the multinomial logit models that are being considered fulfill the definition of a multivariate generalized linear model. Thus multinomial logit models of the form $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W_{ij}\mathbf{u}_i$, where the random effects $\mathbf{u}_i$ satisfy Stage

2 of the definition above, can be considered as MGLMMs. Such a result is advantageous for a number of reasons. As MGLMMs are extensions of multivariate generalized linear models, the score and information functions have known forms which can be utilized in the maximum likelihood estimation routines. The MGLMM framework also provides a unified approach to modeling for the class of multinomial models. Maximum likelihood algorithms can be defined for the general MGLMM and then modified appropriately for the link function, response function, and random effects structure under consideration. For these reasons we present the fitting algorithms in Sections 3.3, 3.5, and 4.2 in terms of a MGLMM for multinomial response data.

3.2.2 Motivation of the Multinomial Logit Models

We have seen in the previous section that multinomial logit random effects models are special cases of MGLMMs. One can also motivate these models by assuming that the true response is an underlying latent variable that follows a linear mixed model. This approach, using a linear fixed effects model, has been used to motivate the fixed effects versions of the baseline-category logit model, the continuation-ratio logit model, and the cumulative logit model. Thus, in the sections below, we begin by motivating the fixed effects models and then extend the motivations to include random effects. The adjacent-category logit model, however, lacks a meaningful motivation based on an underlying latent variable. We present a latent variable motivation that leads to the adjacent-category logit model, but, admittedly, lacks the interpretability of the other motivations.

Baseline-Category logit model

The baseline-category logit model can be motivated from the consideration of an underlying latent variable by using the principle of maximum random utility. The concept of random utility arose out of psychological research by Thurstone (1927) and has been applied in many areas such as consumer theory, transportation theory, and behavioral theory (see, e.g., Ben-Akiva and Lerman 1985). The basic theory is that a consumer who is faced with a finite set of choices, will select the option that provides him or her the maximum use or utility. Since it is impossible to predict the chosen alternative for all consumers due to, for example, possible misperceptions of choices by individuals, Thurstone (1927) considered the true utilities of the possible choices as random variables. Thus the probability that a particular alternative is chosen is defined as the probability that it has the greatest utility among all possible choices.

We first consider the motivation of the fixed effects baseline-category logit model, in which, for convenience, we suppress the subscripts i and j. For a nominal response variable Y, it is assumed that a latent variable Y_r^* is associated with the rth category or choice, $r = 1, ..., R = q+1$. Y_r^* can be thought of as a measure of the utility of the rth category. Proceeding as in Fahrmeir and Tutz (1994, p. 70), let Y_r^* be denoted as

$$Y_r^* = U_r + \epsilon_r, \tag{3.2}$$

where $U_r = \alpha_r + \mathbf{x}'\mathcal{B}_r = \mathbf{z}'\boldsymbol{\beta}_r^*$ is the unobserved utility for the rth alternative and $\epsilon_1, ..., \epsilon_R$ are random variables with some continuous distribution function F. It is assumed that the unobserved utility U_r depends on a vector of covariates $\mathbf{z}' = (1, \mathbf{x}')$ with corresponding parameter vector $\boldsymbol{\beta}_r^{*'} = (\alpha_r, \mathcal{B}_r')$. The principle of maximum random utility assumes that choice r will be chosen if it provides the maximum perceived utility for the consumer. That is, the observed response Y takes on the value r according to

$$Y = r \Leftrightarrow Y_r^* = \max_{l=1,...,R} Y_l^*. \tag{3.3}$$

Thus the response category r is chosen if the underlying latent variable Y_r^* has the maximum utility.

Let $f(\epsilon)$ be the density function of ϵ and assume that the $\{\epsilon_r\}$ are independent. Given the relationship between Y and Y_r^* defined in (3.3) and the model definition (3.2) for Y_r^*, the probability that a consumer chooses alternative r is

$$\begin{aligned} P(Y = r) &= P(Y_r^* - Y_1^* \geq 0, ..., Y_r^* - Y_R^* \geq 0) \\ &= P(\epsilon_1 \leq U_r - U_1 + \epsilon_r, ..., \epsilon_k \leq U_r - U_R + \epsilon_r) \\ &= \int_{-\infty}^{\infty} \prod_{s \neq r} F(U_r - U_s + \epsilon)\, f(\epsilon)\, d\epsilon. \end{aligned} \tag{3.4}$$

The baseline-category logit model is motivated by assuming the $\{\epsilon_r\}$ follow the extreme value distribution, whose distribution is defined by $F(x) = \exp(-\exp(-x))$. Under these assumptions, the integrand of (3.4) takes the form

$$\begin{aligned} \prod_{s \neq r} F(U_r - U_s + \epsilon)\, f(\epsilon) &= \prod_{s \neq r} \exp(-e^{-U_r + U_s - \epsilon}) \exp(-\epsilon - e^{-\epsilon}) \\ &= \exp\left[-\epsilon - e^{-\epsilon} \left(\sum_{s=1}^{R} \frac{e^{U_r}}{e^{U_s}} \right) \right]. \end{aligned} \tag{3.5}$$

Letting $\Delta = \log(\sum_{s=1}^{R} \frac{e^{U_r}}{e^{U_s}})$ in (3.5) and defining $\epsilon^* = \epsilon - \Delta$, the integration in (3.4) yields

$$\begin{aligned} \int_{-\infty}^{\infty} \prod_{s \neq r} F(U_r - U_s + \epsilon)\, f(\epsilon)\, d\epsilon &= \int_{-\infty}^{\infty} \exp(-\epsilon - e^{-(\epsilon - \Delta)})\, d\epsilon \\ &= \exp(-\Delta) \int_{-\infty}^{\infty} \exp(-\epsilon^* - e^{-\epsilon^*})\, d\epsilon^* \\ &= \exp(-\Delta). \end{aligned}$$

Thus the probability that the rth alternative is chosen is

$$\begin{aligned} P(Y = r) &= \frac{\exp(U_r)}{\sum\limits_{s=1}^{R} \exp(U_s)} \\ &= \frac{\exp(U_r - U_R)}{1 + \sum\limits_{s=1}^{q} \exp(U_s - U_R)}. \end{aligned} \tag{3.6}$$

By substituting $\tilde{U}_r = U_r - U_R$ into (3.6), one obtains the baseline-category logit model (2.14), where $\tilde{U}_r = (\alpha_r - \alpha_R) + \mathbf{x}'(\boldsymbol{\beta}^*_r - \boldsymbol{\beta}^*_R) = \mathbf{z}'\boldsymbol{\beta}_r$.

Now consider a nominal response variable Y_{ij} for the jth observation on the ith subject with corresponding latent response Y^*_{ijr} for the rth alternative. The general baseline-category logit random effects model is motivated by assuming that the rth unobserved latent response follows the mixed linear model

$$Y^*_{ijr} = U_{ijr} + \epsilon_{ijr} = \mathbf{z}'_{ij}\boldsymbol{\beta}^*_r + \mathbf{w}'_{ij}\mathbf{u}^*_{ir} + \epsilon_{ijr}, \tag{3.7}$$

where U_{ijr} now includes cluster- and category-specific random effects $\mathbf{u}^*_{ir}$. As before, the $\{\epsilon_{ijr}\}$ are independently and identically distributed random variables that follow the extreme value distribution, F. In addition, the distribution of the cluster-specific random effects $\mathbf{u}^{*'}_i = (\mathbf{u}^{*'}_{i1}, \cdots, \mathbf{u}^{*'}_{iR})$ is assumed to be multivariate normal with mean $\mathbf{0}$ and covariance Σ^*, where the $\{\mathbf{u}^*_i\}$ are distributed independently of the $\{\epsilon_{ijr}\}$. Note that imposing a multivariate distribution on the vector of cluster-specific random effects $\mathbf{u}^*_i$ allows the category-specific random effects to be correlated.

As in Stage 1 of the definition of the multivariate generalized linear mixed model given in Section 3.2.1, motivation proceeds by conditioning on the random effects $\mathbf{u}^*_i$. That is, given the unobserved random effects $\mathbf{u}^*_i$, the nominal response Y_{ij} takes the value r according to

$$Y_{ij} = r \mid \mathbf{u}^*_i \Leftrightarrow Y^*_{ijr} = \max_{l=1,\ldots,R} Y^*_{ijl}. \tag{3.8}$$

Using the same approach as for the fixed effects model, the probability that $Y_{ij} = r$ given the random effects $\mathbf{u}^*_i$ is

$$P(Y_{ij} = r \mid \mathbf{u}^*_i) = \frac{\exp(U_{ijr} - U_{ijR})}{1 + \sum_{s=1}^{q} \exp(U_{ijs} - U_{ijR})}.$$

By letting $\boldsymbol{\beta}_r = \boldsymbol{\beta}_r^* - \boldsymbol{\beta}_R^*$ and $\mathbf{u}_{ir} = \mathbf{u}_{ir}^* - \mathbf{u}_{iR}^*$, the model takes on the usual form

$$P(Y_{ij} = r \mid \mathbf{u}_i) = \frac{\exp(\eta_{ijr})}{1 + \sum_{s=1}^{q} \exp(\eta_{ijs})}, \tag{3.9}$$

where $\eta_{ijr} = z_{ij}^{'}\boldsymbol{\beta}_r + w_{ij}^{'}\mathbf{u}_{ir}$ and $\mathbf{u}_i^{'} = (\mathbf{u}_{i1}, \cdots, \mathbf{u}_{iq})$, with $\mathbf{u}_i$ distributed as a multivariate normal with mean $\mathbf{0}$ and covariance matrix Σ. In matrix notation, the random effects baseline-category logit model takes the form $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W_{ij}\mathbf{u}_i$, with Z_{ij} and $\boldsymbol{\beta}$ defined as in 2.3.1 and W_{ij} defined according to the random effects structure. For example, to allow each threshold to be random, W_{ij} would have the form

$$W_{ij} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix},$$

and $\mathbf{u}_i = (u_{1i}, \cdots, u_{iq})$. Since the thresholds are not ordered in the baseline-category logit model, there are no complications encountered in the estimation procedure when they are allowed to vary individually. Though the motivation in this section included only covariates related to the subject, $\mathbf{z}_{ij}$, inclusion of covariates constant across subjects and specific to the alternatives is straightforward as shown in Section 2.4.1.

Continuation-Ratio logit model

As noted in Section 2.3.4, the continuation-ratio logit model is useful for modeling data in which the ordering of the responses is due to a sequential mechanism. Recall that in the example of that section, tonsils could not be categorized as "Greatly enlarged" unless they had passed through the "Present but not enlarged" and "Enlarged" categories. One method of motivating this model is to consider an underlying latent response which follows a similar sequential process (Fahrmeir and Tutz 1994, p. 85). Again considering the fixed effects model first and suppressing subscripts, let $U_1, \cdots, U_q$ denote q latent variables defined by the linear model $U_r = -\mathbf{x}'\boldsymbol{\gamma} + \epsilon_r$, where

ϵ_r has distribution function F. Considering the response mechanism as sequential, the ordinal response Y starts with the value one according to

$$Y = 1 \Leftrightarrow U_1 \leq \alpha_1,$$

where α_1 denotes a threshold parameter. If U_1 exceeds α_1, the response Y proceeds to category two where

$$Y = 2 \text{ given } Y \geq 2 \Leftrightarrow U_2 \leq \alpha_2.$$

For the rth category, the relation is given by

$$Y = r \text{ given } Y \geq r \Leftrightarrow U_r \leq \alpha_r.$$

The sequential process continues until U_r does not exceed the threshold α_r, at which point the response r is observed.

By replacing U_r with its linear model representation and by considering the equivalent representation of the sequential process

$$Y > r \text{ given } Y \geq r \Leftrightarrow U_r > \alpha_r, \tag{3.10}$$

the probability that $Y = r$ given that $Y \geq r$ is given by

$$P(Y = r \mid Y \geq r) = F(\alpha_r + \mathbf{x}'\gamma), \quad r = 1, \cdots, R, \tag{3.11}$$

where $\alpha_R = \infty$. The continuation-ratio logit model is obtained by taking F to be the logistic distribution, $F(x) = 1/(1 + \exp(-x))$. Denoting $\eta_r = \alpha_r + \mathbf{x}'\gamma$, the unconditional probability that the ordinal response Y takes the value r, $r = 1, \cdots, R$, is given by

$$\begin{aligned} P(Y=r) &= P(Y=r \mid Y \geq r)\ P(Y \geq r) \\ &= F(\eta_r) \prod_{l=1}^{r-1} [1 - F(\eta_r)] \\ &= \frac{1}{1+\exp(-\eta_r)} \prod_{l=1}^{r-1} \left[1 - \frac{1}{1+\exp(-\eta_l)}\right], \end{aligned} \tag{3.12}$$

where $\prod_{r=1}^{0}\{\cdot\} = 1$.

Extending the continuation-ratio logit model to include random effects is straightforward in the motivation given above. Reintroducing subscripts, let the R latent variables for jth observation and the ith subject be denoted by $U_{ij1}, \cdots, U_{ijR}$. To incorporate random effects, one now assumes that $U_{ijr} = -\mathbf{x}_{ij}'\boldsymbol{\gamma} - \mathbf{w}_{ij}'\mathbf{u}_i + \epsilon_{ijr}$, where $\mathbf{u}_i \sim MVN(\mathbf{0}, \Sigma)$ is a vector of cluster-specific random effects, with the $\{\mathbf{u}_i\}$ distributed independently of the $\{\epsilon_{ijr}\}$. The sequential mechanism defined in (3.10) is now assumed to hold conditionally on the random effects $\mathbf{u}_i$. That is

$$\{Y > r \mid \mathbf{u}_i\} \ \text{ given } \ \{Y \geq r \mid \mathbf{u}_i\} \ \Leftrightarrow \ U_r > \alpha_r.$$

Then the conditional probability that $Y_{ij} = r$ given $\mathbf{u}_i$ takes the same form as (3.12) with linear predictor $\eta_{ijr} = \alpha_r + \mathbf{x}_{ij}'\boldsymbol{\gamma} + \mathbf{w}_{ij}'\mathbf{u}_i = \mathbf{z}_{ij}'\boldsymbol{\beta} + \mathbf{w}_{ij}'\mathbf{u}_i$. In terms of the design matrices, the continuation-ratio logit model can be written in familiar form $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W_{ij}\mathbf{u}_i$.

Cumulative logit model

As in the continuation-ratio logit model, the cumulative logit model can be motivated from a threshold approach, which has been shown by numerous authors (see, e.g., Tutz and Hennevogl 1996). For this approach, one assumes that the observed categorical response Y is a categorized version of an underlying latent continuous response Y^*. Observations of Y^* are obtained only through the less precise categorized response Y, which is defined by a set of thresholds $(\alpha_1, \cdots, \alpha_q)$. That is

$$Y = r \quad \Leftrightarrow \quad \alpha_{r-1} < Y^* \leq \alpha_r, \quad r = 1, \cdots, R, \tag{3.13}$$

where $\alpha_0 = -\infty < \alpha_1 < \cdots < \alpha_R = \infty$. Since the $\{\alpha_r\}$ divide the continuous response Y^* into categories, they are often referred to as cut-points. To motivate the cumulative logit model, it is assumed that the latent response Y^* follows a linear model $Y^* = -\mathbf{x}'\boldsymbol{\gamma} + \epsilon$, where $\mathbf{x}$ is an observed covariate vector with corresponding parameter vector $\boldsymbol{\gamma}$, and ϵ follows the distribution function F.

For the general cumulative model, the probability of observing a response $Y \leq r$ is given from (3.13) by

$$P(Y \leq r) = F(\alpha_r + \mathbf{x}'\boldsymbol{\gamma}), \quad r = 1, \cdots, q.$$

A variety of cumulative link models can be defined by specifying F, such as the cumulative probit model when F is the standard normal distribution. To obtain the cumulative logistic model, F is taken to be the logistic distribution function. Hence, the probability that Y takes on the value r is

$$P(Y = r) = \frac{1}{1 + \exp(-\eta_r)} - \frac{1}{1 + \exp(-\eta_{r-1})}, \quad r = 1, ..., q, \tag{3.14}$$

where $\eta_r = \alpha_r + \mathbf{x}'\boldsymbol{\gamma}$ and $\alpha_0 = -\infty$. We note that as a special case of the cumulative logit motivation, binary logistic models can also be motivated using the threshold approach, where the underlying response Y^* is categorized by a single threshold α.

To allow for random effects, the underlying latent variable Y_{ij}^* is assumed to follow a linear model with both fixed and random components. That is, $Y_{ij}^* = -\mathbf{x}_{ij}'\boldsymbol{\gamma} - \mathbf{w}_{ij}'\mathbf{u}_i + \epsilon_{ij}$, with $\mathbf{u}_i \sim MVN(\mathbf{0}, \Sigma)$, $i = 1, \cdots, n$, distributed independently of the $\{\epsilon_{ij}\}$. The relationship between the observed response Y_{ij} and the latent response Y_{ij}^*

is now defined conditionally, such that

$$Y = r \mid \mathbf{u}_i \quad \Leftrightarrow \quad \alpha_{r-1} < Y^* \leq \alpha_r, \quad r = 1, \cdots, R. \tag{3.15}$$

The response probability (3.14) is now defined conditionally on $\mathbf{u}_i$, which follows directly from (3.15). The resulting linear predictor is $\eta_{ijr} = \mathbf{z}_{ij}^{'}\boldsymbol{\beta} + \mathbf{w}_{ij}\mathbf{u}_i$. In Section 3.7 we will consider the motivation of the extended threshold model of Tutz and Hennevogl (1996) which allows each threshold to vary individually.

Adjacent-Category logit model

The motivations of the previous three models have been based on realistic mechanisms for relating the latent response to the observed response. For example, the baseline-category logit model was based on the psychological principle of maximum random utility. For the adjacent-category logit model, however, a meaningful motivation has not been established. Recall that in this model a common association parameter $\boldsymbol{\beta}$ is assumed to hold for all logits constructed from adjacent response categories. As the model is defined for ordinal responses, one might consider a motivation based on a threshold approach. However, the threshold approach is not appropriate since the thresholds in the adjacent-category logit model need not be ordered. The sequential process of the ordinal continuation-ratio logit model does not apply as well. We outline below a motivation for the adjacent-category logit model that parallels that of the baseline-category logit model. The motivation does lead to the adjacent-category logit model, however it lacks the realistic interpretation that the other motivations possess.

Though the model is defined for ordinal responses, the adjacent-category logit model has response probabilities that are very similar to that of the baseline-category logit model (compare (2.17) with (2.14)). Thus one approach for motivating the adjacent-category logit model is that based on a modified principle of maximum random utility. Recall that the principle of maximum random utility assumes that a subject, given a set of R nominal choices, will choose the alternative that maximizes his or her utility. For the adjacent-category motivation, one could assume that a subject chooses from a set of R ordered alternatives, where the utility $Y_r^\star$ for the rth ordered alternative is defined by the model

$$Y_r^\star = U_r + \epsilon_r = r\,\alpha_r + r\,\mathbf{x}'\boldsymbol{\mathcal{B}} + \epsilon_r, \quad r = 1, \cdots, R. \tag{3.16}$$

Note that (3.16) differs from (3.2) in two respects. First (3.16) has a modified linear predictor in that it is scaled by the ordered category choice, and (3.16) has a common parameter $\boldsymbol{\mathcal{B}}$ across all utilities. The relationship between the observed ordinal response Y and the latent responses $Y_r^\star$, $r = 1, \cdots, R$ is given by

$$Y = r \Leftrightarrow Y_r^\star = \max_{l=1,\ldots,R} Y_l^\star. \tag{3.17}$$

By assuming that ϵ_r in (3.16) follows the extreme value distribution and then calculating the probability that $Y = r$ according to (3.17), the probability of the rth ordered alternative can be shown to be

$$P(Y = r) = \frac{\exp(\eta_r)}{1 + \sum_{l=1}^{q} \exp(\eta_l)}, \quad r = 1, \ldots q, \tag{3.18}$$

where $\eta_r = (r - R)\,\mathbf{z}'\boldsymbol{\beta}$ with $\mathbf{z}' = (1, \mathbf{x}')$ and $\boldsymbol{\beta}' = (\alpha_1, \cdots, \alpha_q, \boldsymbol{\mathcal{B}}')$. The inclusion of random effects is straightforward, following the same conditioning arguments found in the baseline-category logit motivation.

One criticism of the motivation given above is that the principle of maximum random utility was conceived for, and intuitively makes sense for situations in which the possible alternatives are nominal. However, one could argue that the use of the principle of maximum random utility in the ordinal setting could be appropriate as well. Consider a situation in which a group of subjects is asked to rate their job satisfaction on an ordinal scale from "Very dissatisfied" to "Very satisfied". The response given by a subject is the one that most epitomizes or "maximizes" their feelings. A second criticism is that the form of the utility model given in (3.16) is contrived. It is obviously chosen so that the desired form of the adjacent-category logit model is obtained. However, if one wished to motivate a baseline-category logit model that allowed, say, alternative-specific covariates which varied across choosers, one would simply modify (3.2) appropriately so that the desired model was obtained as well. Thus we contend that the given motivation can be appropriate for the adjacent-category logit model.

3.3 Maximum Likelihood Estimation

From the definition given in Section 3.2.1 and the independence of observations between clusters, the form of the marginal likelihood for the multivariate generalized linear mixed model with response vector $\mathbf{y}_{ij}$ is immediately given by

$$L(\boldsymbol{\beta}, G) = \prod_{i=1}^{n} \int \cdots \int \left[\prod_{j=1}^{T_i} f(\mathbf{y}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i) \right] \, dG(\mathbf{u}_i).$$

For the multinomial random effects models considered in this chapter, $f(\cdot \mid \boldsymbol{\beta}; u_i)$ is the multinomial distribution and $G(\mathbf{u}_i)$ is the multivariate normal distribution with mean $\mathbf{0}$ and covariance Σ. Thus the marginal likelihood for the general multinomial random effects model is

$$L(\boldsymbol{\beta}, \Sigma) = \prod_{i=1}^{n} \int \cdots \int \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i) \right] \, (2\pi)^{-m/2} \mid \Sigma \mid^{-1/2} \exp(-\frac{1}{2} \mathbf{u}_i' \Sigma^{-1} \mathbf{u}_i) \, d\mathbf{u}_i, \tag{3.19}$$

where $\mathbf{u}' = (\mathbf{u}_1', \cdots, \mathbf{u}_n')$ and m is the dimension of the random effects vector $\mathbf{u}_i$. Note that the multinomial distribution $f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)$ in (3.19) is defined in terms of the scaled multinomial response $\bar{\mathbf{y}}_{ij} = \mathbf{y_{ij}}/n_{ij}$, where n_{ij} is the multinomial index associated with the response $\mathbf{y}_{ij}$. In many longitudinal datasets with categorical responses, the multinomial index for each observation will be one. We will, however, define the general multinomial random effects model in terms of $\bar{\mathbf{y}}_{ij}$, with the understanding that n_{ij} may be one for all subjects and observations.

In this section we present methods for finding maximum likelihood estimates of $\boldsymbol{\beta}$ and Σ for the general multinomial random effects model. Finding maximum likelihood estimates entails maximizing the marginal likelihood (3.19). To do this, however, one must first evaluate an intractable m-dimensional integral for each cluster. Thus methods for finding maximum likelihood estimates must incorporate both a maximization routine and an algorithm for approximating integrals. Technically, such methods provide only approximate maximum likelihood estimates since the intractable integrals are only approximated. However, we also propose in Section 3.5 a routine which fits linear mixed models to a Taylor series approximation of the link function to obtain approximate maximum likelihood estimates. Thus we refer to the methods in this section as maximum likelihood, and reserve the term approximate for the method in Section 3.5. In this section we discuss two such maximum likelihood methods. The first directly maximizes (3.19) using adaptive Gauss-Hermite quadrature to approximate the intractable integrals. The second uses the EM algorithm to indirectly maximize (3.19) using Monte Carlo techniques to evaluate the integrals.

3.3.1 Maximum Likelihood Algorithms

There has been a considerable amount of recent research focused on accurate and efficient methods for obtaining maximum likelihood estimates in generalized linear mixed models (Zeger and Karim 1991; McCulloch 1997; Booth and Hobert 1999). Generally, these methods can be categorized by whether they directly or indirectly maximize the marginal (log) likelihood, and by whether they use deterministic or random sampling for the numerical integration. For direct maximization, the marginal (log) likelihood is obtained through numerical integration and then directly maximized by, for example, using Fisher scoring. Indirect maximization is accomplished through the EM algorithm, a powerful iterative method for obtaining maximum likelihood estimators in incomplete data situations. As in the direct maximization method, the E-step in the EM algorithm for generalized linear mixed models contains intractable integrals and so numerical integration is required. Many authors have suggested using Gauss-Hermite quadrature or Monte Carlo integrations techniques to evaluate the intractable integrals in the direct and indirect maximization methods (Fahrmeir and Tutz 1994, Chap. 7; McCulloch 1997; Booth and Hobert 1999).

Gauss-Hermite quadrature is an example of a deterministic sampling approach to numerical integration. In general, quadrature rules are based on re-expressing a regular function $f(x)$ as the product of a known weight function $w(x)$ and another function $h(x)$ such that $f(x) = w(x)\,h(x)$. Then, for predetermined nodes ς_l and weights ϖ_l, an integral can be approximated by a discrete summation,

$$\int f(x)dx = \int w(x)\,h(x)dx \approx \sum_{l=1}^{K} \varpi_l\, h(\varsigma_l), \tag{3.20}$$

where the integration is over the domain of $f(x)$. Gauss-Hermite integration (Stroud and Secrest 1966) is popular in statistics due to the form of the weight function $w(x)$ and the bounds of the integration. Specifically, univariate Gauss-Hermite quadrature approximates integrals over the real line that can be expressed in terms of the weight function $w(x) = \exp(-x^2)$. Thus for univariate Gauss-Hermite quadrature, approximation (3.20) takes the form

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{+\infty} \exp(-x^2)\,h(x)dx \approx \sum_{l=1}^{K} \varpi_l\, h(\varsigma_l), \tag{3.21}$$

where the weights ϖ_l and nodes ς_l are obtained from Stroud and Secrest (1966) or, more conveniently, by using an algorithm proposed by Golub (1973). If $h(x)$ is a polynomial of degree less than or equal to $2K-1$, the approximation given by (3.21) is exact. Thus (3.21) can be made arbitrarily accurate by increasing the number of nodes K. The Gauss-Hermite quadrature rule (3.21) is easily applied to statistical integration problems in which the normal density $g_N(x;\mu,\sigma^2)$ can be expressed as a weight function, $w(x) = g_N(x;\mu,\sigma^2)$. Such integrals can be approximated using (3.21) by substituting ϖ_l and ς_l with $\pi^{-1/2}\,\varpi_l$ and $\sqrt{2}\,\sigma\,\varsigma_l + \mu$, respectively.

The Gauss-Hermite rule (3.21) can be extended to m-dimensional integrals if the weight function is of the form $w(\mathbf{x}) = \exp(-\sum_{i=1}^{m} x_i^2)$. If the weight function is the multivariate normal distribution $g_{\mathtt{MVN}}(\mathbf{x};\boldsymbol{\mu},\Sigma)$, the m-dimensional Gauss Hermite rule takes the form

$$\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty} g_{\mathtt{MVN}}(\mathbf{x};\boldsymbol{\mu},\Sigma)\,h(\mathbf{x})d\mathbf{x} \approx \sum_{l_1=1}^{K}\cdots\sum_{l_m=1}^{K}\pi^{-m/2}\,\varpi_{\mathbf{l}}\,h(\varsigma_{\mathbf{l}}^{\star}), \tag{3.22}$$

where $\varpi_{\mathbf{l}} = \prod_{i=1}^{m}\varpi_{l_i}$, $\varsigma_{\mathbf{l}}^{\star} = \sqrt{2}\,\Sigma^{1/2}\,\varsigma_{\mathbf{l}} + \boldsymbol{\mu}$, $\varsigma_{\mathbf{l}}^{'} = (\varsigma_{l_1},\cdots,\varsigma_{l_m})$, $\Sigma^{1/2}$ denotes the left Cholesky (lower triangular) square root of Σ, and $\{\varpi_{l_i}\}$ and $\{\varsigma_{l_i}\}$, $i = 1,\cdots,m$, are the weights and nodes from the univariate Gauss-Hermite rule. Note that the summation in (3.22) is over K^m nodes so that the number of nodes increases exponentially with the number of dimensions m. Thus multivariate Gauss-Hermite quadrature is currently only computationally feasible for integral dimensions of up to five or six. It should be noted, however, that K should be chosen large enough to ensure accurate approximation of the integrals.

In contrast to Gauss-Hermite quadrature, Monte Carlo methods use randomly sampled nodes to approximate integrals. In Monte Carlo integration techniques the intractable integrals are viewed as expectations with respect to some density function. For example, consider the integral

$$I = \int_{-\infty}^{+\infty} h(x)\,g(x)dx, \tag{3.23}$$

where $h(x)$ is a continuous function and $g(x)$ is a density. A simple Monte Carlo method for calculating (3.23) consists of approximating I by

$$\hat{I} = \frac{1}{K}\sum_{l=1}^{K} h(\varsigma_l),$$

where ς_l, $i = 1, \cdots, K$ are independently and identically distributed samples from the density $g(x)$. By the Law of Large Numbers, $\hat{I}$ converges almost certainly to I (Tanner 1996, p. 51). More sophisticated and efficient approximations based on Monte Carlo methods, such as importance sampling and rejection sampling (see, e.g., Tanner 1996, p. 54), can also be utilized. Extensions to multivariate integrals is straightforward where samples are then drawn from a multivariate candidate distribution. Computationally, Monte Carlo methods are more attractive than Gauss-Hermite methods as the number of draws does not increase exponentially with the dimension of the integrals. It is also possible to assess the error in the integral approximations, as pointed out by Booth and Hobert (1999) and discussed shortly. The assessment of the error in Gauss-Hermite quadrature is extremely complicated, requiring evaluation of the $2K$th derivative of the integrand being approximated.

The approaches used to fit multivariate generalized linear mixed models have generally been modeled after those used for generalized linear mixed models. For single random effects models, a Gauss-Hermite EM algorithm was used by Jansen (1990), while Ezzet and Whitehead (1991) utilized a Gauss-Hermite Newton-Raphson algorithm. Both Hedeker and Gibbons (1994) and Tutz and Hennevogl (1996) proposed algorithms for fitting general ordinal regression models that allowed multiple random effects. Hedeker and Gibbons (1994) and Hedeker (2000) considered a direct maximization approach using multivariate Gauss-Hermite quadrature to approximate the integrals in conjunction with Fisher's method of scoring for maximization. They, however, used a modified multivariate Gauss-Hermite rule as compared to that given in (3.22). They orthogonally transformed the response model so that the weight function was the multivariate standard normal. Specifically, assume that the vector of random effects $\mathbf{u}_i \sim MVN(\mathbf{0}, \Sigma)$ and let $\Sigma = \Sigma^{1/2}\, \Sigma^{1/2'}$ where $\Sigma^{1/2}$ is the left Cholesky factor of Σ. Then the linear predictor $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W_{ij}\mathbf{u}_i$ can be transformed to $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W_{ij}\, \Sigma^{1/2}\, \mathbf{a}_i$ where $\mathbf{a}_i$ has the multivariate standard normal distribution.

Using some matrix algebra, the more common linear form can be obtained

$$\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W^*_{ij}\boldsymbol{\theta}, \tag{3.24}$$

where $W^*_{ij} = \mathbf{a}^{'}_i \otimes W_{ij}$, $\boldsymbol{\theta} = \text{vech}(\Sigma^{1/2})$, and $\otimes$ denotes the Kronecker product. Note that vech(M) denotes a column vector with the unique stacked rows of the upper triangle of M. That is, for example,

$$\text{vech}(M) = \text{vech}\begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix} = \begin{bmatrix} m_{11} \\ m_{12} \\ m_{22} \end{bmatrix}.$$

Maximum likelihood estimates are then obtained for the fixed effects parameter vector $\boldsymbol{\beta}$ and the unique elements $\boldsymbol{\theta}$ of the Cholesky factor of Σ. A consequence of this approach is that the weight function becomes the multivariate standard normal distribution. Thus the univariate Gauss-Hermite nodes no longer need to be transformed as in (3.22). Another advantage is that estimation of the Cholesky elements $\boldsymbol{\theta}$ instead of the covariance elements of Σ is often more stable when the true variance elements are near zero. A disadvantage is that it is often difficult, and sometimes impossible to determine the relationship between the parameters in Σ and the parameters in $\boldsymbol{\theta}$ when constraints are placed on the elements of Σ.

Tutz and Hennevogl (1996) considered an indirect maximization approach for fitting their general ordinal random effects model. Using the transformed linear predictor (3.24), they outlined both a multivariate Gauss-Hermite EM algorithm and a Monte Carlo EM algorithm. In both algorithms they replaced the intractable integrals in the E-step by numerical approximations. Due to the transformation (3.24), the Monte Carlo approximation required sampling from the multivariate standard normal distribution. An important issue in both Monte Carlo integration and Gauss-Hermite quadrature is determining the number of samples, in the former, or nodes, in the latter, that are required to adequately approximate the integrals. For models with a single random effect, Tutz and Hennevogl (1996) suggested that only 10 to 20 samples were needed for the Monte Carlo integration and only 8 to 10 quadrature points for Gauss-Hermite integration. When they considered a four-dimensional random effect, they only recommended Monte Carlo integration with 20 to 30 samples as they reported extremely high computational time for the Gauss-Hermite EM algorithm. For their Gauss-Hermite Fisher scoring algorithm, Hedeker and Gibbons (1994) suggested that the number of quadrature points could actually be decreased as the number of random effects was increased. For example, they suggested that as few as three points per dimension for a five-dimensional random effects model would be adequate for approximating the integrals. From our experience and others (Pinheiro and Bates 1995; Agresti and Hartzel 1999), these recommendations will not provide adequate approximations of the integrals. For the nonlinear mixed effects logistic model with a single random effect, Pinheiro and Bates (1995) reported that over 100 quadrature points were required to obtain the correct maximum likelihood estimates. Agresti and Hartzel (1999) reported similar results for the logistic-normal model. In Section 3.6 we illustrate these findings for models with ordinal responses. The accuracy of using only 20 to 30 Monte Carlo samples is also questionable, since no attempt was made to assess the Monte Carlo error in the approximations.

We now propose two algorithms for approximating and maximizing the marginal likelihood (3.19) of the general multinomial random effects model. The first method utilizes adaptive Gauss-Hermite quadrature (Liu and Pierce 1994; Pinheiro and Bates 1995) to numerically approximate the intractable integrals, and then directly maximizes the marginal likelihood with a quasi-Newton algorithm. Thus we follow the direct maximization approach of Hedeker and Gibbons (1994) and Hedeker (2000). Note, however, that our algorithm differs with respect to their algorithm in two ways. First, we utilize adaptive quadrature which, as noted above, often requires substantially fewer points for obtaining the same decimal accuracy as Gauss-Hermite quadrature. Secondly, we utilize a quasi-Newton algorithm in place of their Fisher scoring algorithm. As noted in Chapter 2, the main difference between the Fisher scoring algorithm and the Newton-Raphson algorithm is that the expected information matrix is used in the former, while the observed information matrix is used in the latter. Thus the Fisher scoring algorithm has the advantage of using an information matrix that is always positive definite. For random effects models, however, we feel that the observed information matrix is easier to calculate. In general, the Fisher scoring algorithm and the Newton-Raphson algorithm will perform similarly for a given situation.

In our second proposed algorithm a Monte Carlo EM algorithm is used to indirectly maximize (3.19). Specifically, we apply the Monte Carlo EM algorithm of Booth and Hobert (1999) which allows for the assessment of the Monte Carlo error in the integral approximations. This approach is quite different from that of Tutz and Hennevogl (1996) who did not evaluate the error in the Monte Carlo integration. In addition, the number of Monte Carlo samples was fixed at the start of the algorithm. In our approach, the sample size can be increased after each iteration of the EM algorithm. Since the EM algorithm is inherently slow, we recommend its use in situations where the integral dimensions are extremely high, for which the adaptive quadrature algorithm becomes too computationally burdensome. For most applications with low to moderate numbers of random effects, such as those considered in Section 3.6, the adaptive quadrature algorithm will provide the most efficient means for finding maximum likelihood estimates.

3.3.2 Quasi-Newton, Adaptive Gauss-Hermite Quadrature Algorithm

Denoting the multivariate normal density by $g_{\mathrm{MVN}}(\mathbf{x};\boldsymbol{\mu},\Sigma)$, the ith set of integrals, $i=1,\cdots,n$, in (3.19) is

$$L_i(\boldsymbol{\beta},\Sigma)=\int\cdots\int\left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij}\mid\boldsymbol{\beta};\mathbf{u}_i)\right] g_{\mathrm{MVN}}(\mathbf{u}_i;\mathbf{0},\Sigma)\,d\,\mathbf{u}_i. \tag{3.25}$$

The multivariate Gauss-Hermite approximation to (3.25) is given by (3.22) with $h(\cdot)$ replaced with $\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij}\mid\boldsymbol{\beta};\mathbf{u}_i)$. Such an approximation centers and scales the original Gauss-Hermite nodes to have the same mean and variance as the multivariate normal density, $g_{\mathrm{MVN}}(\mathbf{u}_i;\mathbf{0},\Sigma)$. Liu and Pierce (1994), for approximating a single integral, and Pinheiro and Bates (1995), for approximating multiple integrals in nonlinear mixed effects models, discussed a modified Gauss-Hermite rule which Pinheiro and Bates (1995) called adaptive Gauss-Hermite quadrature. Both Liu and Pierce (1994) and Pinheiro and Bates (1995) noted that scaling the nodes about the mean and variance of just $g_{\mathrm{MVN}}(\mathbf{u}_i;\mathbf{0},\Sigma)$ was inefficient as it failed to take into account the impact of $h(\cdot)$ on the shape of the integrals to be approximated. Instead, they recommended the following approach.

Adaptive Gauss-Hermite quadrature

Begin by letting

$$h(\mathbf{u}_i)=\left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij}\mid\boldsymbol{\beta};\mathbf{u}_i)\right] g_{\mathrm{MVN}}(\mathbf{u}_i;\mathbf{0},\Sigma), \tag{3.26}$$

so that (3.25) becomes

$$\int\cdots\int h(\mathbf{u}_i)\,d\,\mathbf{u}_i. \tag{3.27}$$

With adaptive Gauss-Hermite quadrature, one uses the mode $\boldsymbol{\mu}_i^*$ of $h(\mathbf{u}_i)$, and the curvature Σ_i^* around the mode of $h(\mathbf{u}_i)$ to center and scale the original Gauss-Hermite nodes. To utilize Gauss-Hermite quadrature, (3.27) must be written in the appropriate form which requires the weight function $\exp(-\sum_{l=1}^{m} u_{il}^2)$. To this end (3.27) is rewritten

$$\int \cdots \int \frac{h(\mathbf{u}_i)}{g_{\mathtt{MVN}}(\mathbf{u}_i; \boldsymbol{\mu}_i^*, \Sigma_i^*)} \, g_{\mathtt{MVN}}(\mathbf{u}_i; \boldsymbol{\mu}_i^*, \Sigma_i^*) \, d\, \mathbf{u}_i, \tag{3.28}$$

where $g_{\mathtt{MVN}}(\mathbf{u}_i; \boldsymbol{\mu}_i^*, \Sigma_i^*)$ is the multivariate normal density with mean $\boldsymbol{\mu}_i^*$ and covariance matrix Σ_i^*. The multivariate Gauss-Hermite rule (3.22) can now be applied to (3.28) with $h(\cdot)$ in (3.22) replaced with

$$\frac{h(\mathbf{u}_i)}{g_{\mathtt{MVN}}(\mathbf{u}_i; \boldsymbol{\mu}_i^*, \Sigma_i^*)}.$$

The adaptive Gauss-Hermite quadrature approximation of (3.25) is then given by

$$L_i(\boldsymbol{\beta}, \Sigma) = \int \cdots \int \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i) \right] \, g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma) \, d\, \mathbf{u}_i \approx$$
$$\mid \Sigma_i^* \mid^{1/2} 2^{m/2} \sum_{\mathbf{l}_1}^{K} \cdots \sum_{\mathbf{l}_m}^{K} \varpi_{\mathbf{l}} \left[\prod_{j=1}^{T_i} f(\mathbf{y}_{ij} \mid \boldsymbol{\beta}; \boldsymbol{\varsigma}_{i\mathbf{l}}^{\star}) \right] g_{\mathtt{MVN}}(\boldsymbol{\varsigma}_{i\mathbf{l}}^{\star}; \mathbf{0}, \Sigma) \exp(\boldsymbol{\varsigma}_{\mathbf{l}}^{'} \, \boldsymbol{\varsigma}_{\mathbf{l}}), \tag{3.29}$$

where $\varpi_{\mathbf{l}} = \prod_{i=1}^{m} \varpi_{l_i}$, $\boldsymbol{\varsigma}_{i\mathbf{l}}^{\star} = \sqrt{2}\, \Sigma_i^{*1/2} \boldsymbol{\varsigma}_{\mathbf{l}} + \boldsymbol{\mu}_i^*$, $\boldsymbol{\varsigma}_{\mathbf{l}}^{'} = (\varsigma_{l_1}, \cdots, \varsigma_{l_m})$, and $\{\varpi_{l_i}\}$ and $\{\varsigma_{l_i}\}$, $i = 1, \cdots, m$, are the weights and nodes from the univariate Gauss-Hermite rule. Adaptive Gauss-Hermite quadrature can be viewed as a deterministic version of Monte Carlo integration in which the random samples generated from $g_{\mathtt{MVN}}(\mathbf{u}_i; \boldsymbol{\mu}_i^*, \Sigma_i^*)$ are replaced with the fixed values $\{\varsigma_{l_i}^{\star}\}$.

The use of adaptive Gauss-Hermite quadrature can substantially decrease the number of nodes needed to obtain adequate approximations of the intractable integrals (Pinheiro and Bates 1995; Agresti and Hartzel 1999). For a logistic-normal model with a random intercept, Agresti and Hartzel (1999) needed only 9 quadrature points to obtain convergence to four decimal places with adaptive Gauss-Hermite quadrature, as opposed to about 200 for standard Gauss-Hermite quadrature. Similar results are shown in Section 3.6. Adaptive Gauss-Hermite quadrature is computationally more complex than standard Gauss-Hermite quadrature, however. For each subject i, the mode $\boldsymbol{\mu}_i^*$ of $h(\mathbf{u}_i)$ in (3.26), and the curvature Σ_i^* at the mode of $h(\mathbf{u}_i)$

must be calculated. These estimates, based on the current estimates of $\boldsymbol{\beta}$ and vech(Σ), must be recalculated for each iteration of the maximization routine. The mode $\boldsymbol{\mu}_i^*$ can be found using any maximization routine. An estimate of the curvature around the mode can be obtained by inverting the negative of the second derivative matrix of $h(\mathbf{u}_i)$ evaluated at the estimated mode. We use numerical second derivatives to obtain $\hat{\Sigma}_i^*$.

Quasi-Newton algorithm

We incorporate the adaptive Gauss-Hermite approximation into a quasi-Newton algorithm to directly maximize the log of (3.19) for the general multinomial random effects model. Specifically, we utilize the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm which was proposed independently by Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970). The BFGS algorithm was designed for maximizing scalar objective functions and so is well suited for maximum likelihood estimation. Denoting the complete vector of parameters by $\boldsymbol{\Psi}' = (\boldsymbol{\beta}', \text{vech}(\Sigma)')$, the BFGS quasi-Newton algorithm is defined at the $(s+1)$th iteration by

$$\hat{\boldsymbol{\Psi}}^{(s+1)} = \hat{\boldsymbol{\Psi}}^{(s)} + \delta^{(s)}\,\mathcal{H}^{(s)\,-1}\,\mathbf{g}^{(s)}, \tag{3.30}$$

where $\hat{\boldsymbol{\Psi}}^{(s)}$ is the estimate of $\boldsymbol{\Psi}$ at the previous iteration, $\delta^{(s)}$ is a step length between zero and one, and $\mathbf{g}^{(s)}$ and $\mathcal{H}^{(s)}$ denote the gradient vector and Hessian matrix of the log of (3.19), respectively. For our application of the BFGS algorithm, we analytically compute $\mathbf{g}$ by computing analytical first derivatives of the log of (3.19) with respect to $\boldsymbol{\Psi}$ which are given below. Calculation of the Hessian matrix $\mathcal{H}$ is not required as the BFGS algorithm updates this automatically. The step length δ is normally set to one. However, if no improvement is made in the log likelihood value with $\delta = 1$, a δ that improves the value is found by a line search. Convergence is based on both the change in the parameter estimates and the change in the gradient vectors.

We choose to provide analytical first derivatives for the gradient function $\mathbf{g}$ instead of using numerical derivatives. Though the analytical derivatives require additional calculation, they provide increased accuracy over numerical derivatives and are needed for calculation of the observed information matrix upon convergence of the algorithm. Denote the log of (3.19) by $l(\boldsymbol{\beta}, \Sigma)$ and define $L_i(\boldsymbol{\beta}, \Sigma)$ as in (3.25). We first consider the derivative of $l(\boldsymbol{\beta}, \Sigma)$ with respect to $\boldsymbol{\beta}$. Interchanging integrals and derivatives and using the identity

$$\frac{d}{d\boldsymbol{\beta}}\left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)\right] = \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)\right]\left[\sum_{j=1}^{T_i} \frac{d \log f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)}{d\boldsymbol{\beta}}\right], \tag{3.31}$$

one obtains

$$\frac{dl(\boldsymbol{\beta}, \Sigma)}{d\boldsymbol{\beta}} = \sum_{i=1}^{n}\left\{\frac{1}{L_i(\boldsymbol{\beta}, \Sigma)}\right.$$
$$\left.\times \int \cdots \int \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)\right] g_{\texttt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma) \left[\sum_{j=1}^{T_i} \frac{d \log f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)}{d\boldsymbol{\beta}}\right] d\mathbf{u}_i\right\}. \tag{3.32}$$

Denoting the approximation of $L_i(\boldsymbol{\beta}, \Sigma)$ in (3.29) by $\tilde{L}_i$ and using the centered and scaled nodes $\{\varsigma_{i\mathbf{l}}^{\star}\}$ from (3.29), the derivative in (3.32) can be approximated by

$$\frac{dl(\boldsymbol{\beta}, \Sigma)}{d\boldsymbol{\beta}} \approx \sum_{i=1}^{n}\sum_{j=1}^{T_1}\sum_{\mathbf{l}_1}^{K}\cdots\sum_{\mathbf{l}_m}^{K} c_{i\mathbf{l}} \frac{d \log f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \varsigma_{i\mathbf{l}}^{\star})}{d\boldsymbol{\beta}}, \tag{3.33}$$

where

$$c_{i\mathbf{l}} = \frac{1}{\tilde{L}_i}\left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \varsigma_{i\mathbf{l}}^{\star})\right] g_{\texttt{MVN}}(\varsigma_{i\mathbf{l}}^{\star}; \mathbf{0}, \Sigma) \;\; | \, \Sigma_i^{*} \, |^{1/2} \;\; 2^{m/2} \, \varpi_{\mathbf{l}} \, \exp(\varsigma_{\mathbf{l}}^{'} \, \varsigma_{\mathbf{l}}). \tag{3.34}$$

Note that $c_{i\mathbf{l}}$ in (3.34) evaluates to a scalar. Then since $f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \varsigma_{i\mathbf{l}}^{\star})$ is a multivariate generalized linear model, (3.33) can be viewed as a weighted score function with weights $\{c_{i\mathbf{l}}\}$. Using the form of a weighted score function (Tutz and Hennevogl

1996), the approximate derivative (3.33) can be calculated from

$$\frac{dl(\boldsymbol{\beta},\Sigma)}{d\boldsymbol{\beta}} \approx \sum_{i=1}^{n} \sum_{\mathbf{l}_1}^{K} \cdots \sum_{\mathbf{l}_m}^{K} \sum_{j=1}^{T_1} c_{i\mathbf{l}}\, Z'_{ij}\, D_{ij}\, R^{-1}_{\boldsymbol{\pi}_{ij}}\, (\bar{\mathbf{y}}_{ij} - \boldsymbol{\pi}_{ij}), \tag{3.35}$$

where D_{ij}, $R_{\boldsymbol{\pi}_{ij}}$, and Z_{ij} depend on the multinomial link and model, and are given in Chapter 2. Note that (3.35) is evaluated by plugging in the current estimates of $\boldsymbol{\beta}$ and Σ.

We proceed now with the derivative of $l(\boldsymbol{\beta},\Sigma)$ with respect to vech(Σ), the unique elements of Σ. Again interchanging integrals and derivatives and using a similar identity to (3.31),

$$\frac{dl(\boldsymbol{\beta},\Sigma)}{d\,\mathrm{vech}(\Sigma)} = \sum_{i=1}^{n} \Bigg\{ \frac{1}{L_i(\boldsymbol{\beta},\Sigma)} \\ \times \int \cdots \int \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i) \right] g_{\text{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)\, \frac{d \log g_{\text{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)}{d\,\mathrm{vech}(\Sigma)}\, d\mathbf{u}_i \Bigg\}. \tag{3.36}$$

Using the same approach as for (3.33), the derivative (3.36) can be approximated by

$$\frac{dl(\boldsymbol{\beta},\Sigma)}{d\,\mathrm{vech}(\Sigma)} \approx \sum_{i=1}^{n} \sum_{\mathbf{l}_1}^{K} \cdots \sum_{\mathbf{l}_m}^{K} c_{i\mathbf{l}}\, \frac{d \log g_{\text{MVN}}(\boldsymbol{\varsigma}^{\star}_{i\mathbf{l}}; \mathbf{0}, \Sigma)}{d\,\mathrm{vech}(\Sigma)}, \tag{3.37}$$

with $c_{i\mathbf{l}}$ given in (3.34). By substituting into (3.37) the first derivative of the log of a multivariate normal density with respect to the unique elements of Σ, one obtains the approximation,

$$\frac{dl(\boldsymbol{\beta},\Sigma)}{d\,\mathrm{vech}(\Sigma)} \approx \sum_{i=1}^{n} \sum_{\mathbf{l}_1}^{K} \cdots \sum_{\mathbf{l}_m}^{K} c_{i\mathbf{l}} \\ \left[-\Sigma^{-1} + \frac{1}{2}\,\mathrm{diag}(\Sigma^{-1}) + \Sigma^{-1}\, \boldsymbol{\varsigma}^{\star}_{i\mathbf{l}}\, \boldsymbol{\varsigma}^{\star\prime}_{i\mathbf{l}}\, \Sigma^{-1} - \frac{1}{2}\,\mathrm{diag}(\Sigma^{-1}\, \boldsymbol{\varsigma}^{\star}_{i\mathbf{l}}\, \boldsymbol{\varsigma}^{\star\prime}_{i\mathbf{l}}\, \Sigma^{-1}) \right], \tag{3.38}$$

where diag(M) is the matrix M with all off-diagonal elements set to zero. Evaluation of (3.38) is accomplished by plugging in the current estimates of $\boldsymbol{\beta}$ and Σ.

Programming

The matrix programming language Ox (Doornick 1998) was used to program the quasi-Newton, adaptive Gauss-Hermite quadrature algorithm. Ox was designed for programming in matrices and computing matrix calculations, and thus is especially suited for statistical programming. To take advantage of Ox's matrix capabilities, and to avoid the multiple summations found in, for example, (3.38), we expanded the data vectors to match the total number of quadrature points K^m (Hinde 1982). Thus for observation j of subject i, $\mathbf{y}_{ij}^* = \mathbf{1}_{K^m} \otimes \mathbf{y}_{ij}$ and $\mathbf{x}_{ij}^* = \mathbf{1}_{K^m} \otimes \mathbf{x}_{ij}$, where $\mathbf{1}_{K^m}$ is K^m by one column vector of ones. One can then create expanded design matrices Z_{ij}^* and define the random effects design matrix to be $W_{ij}^* = [\varsigma_{i1}^{\star'}]$ which contains the K^m stacked row vectors of the scaled and centered nodes for the ith subject. One must be careful, however, as the matrices can become very large. For example, the dimension of Z_{ij}^* would be $(R-1) * K^m$ by $(R-1)+p$.

Naive starting values for the fixed effects parameters can be obtained from fitting the model without the random effects. For the covariance matrix Σ, one can start with the identity matrix as naive initial estimates. Alternatively, one could use the final estimates from an approximate maximum likelihood algorithm, as will be described in Section 3.5. These estimates are usually close to the true maximum likelihood estimates, and thus reduce the time required for the algorithm to converge. We have also had success using initial estimates obtained from fitting the quasi-Newton algorithm with a smaller number, say five in each dimension, of quadrature points. This is often useful for fitting a model with a large number of random effects, where starting from naive estimates with a large number of quadrature points in each dimension would require many more likelihood evaluations to obtain the maximum. In terms of the number of quadrature points to use, there is no golden number that will ensure adequate approximation of the integrals for every dataset. We recommend upwards of 15 per dimension. However, one should always try additional runs beyond the chosen number to ensure the estimates have stabilized. From our experience, higher numbers of quadrature points are needed to obtain convergence in the standard errors than in the parameter estimates, as was also noted for binary random effects models by Agresti and Hartzel (1999).

3.3.3 Monte Carlo EM Algorithm

The algorithm of the previous section is very efficient when the number of random effects is less than five or six. Beyond this point, however, the computational burden of evaluating K^m points for each observation within each iteration becomes too great. Alternatively, one can use Monte Carlo methods to approximate the intractable integrals. Monte Carlo methods do not experience the curse of dimensionality problems that plagues Gauss-Hermite integration, as the number of samples K does not increase exponentially with the number of random effects (we note that the number of samples needed to approximate integrals does increase with m, but generally not at an exponential rate). Monte Carlo methods can also be formulated so that one can assess the error in the integral approximations (Booth and Hobert 1999). The Monte Carlo EM algorithm proposed by Tutz and Hennevogl (1996) lacked any assessment of Monte Carlo error. This, compounded with the few samples drawn to approximate each integral (maximum of 30) makes estimates from that algorithm suspect.

Booth and Hobert (1999) proposed an automated Monte Carlo EM algorithm for fitting generalized linear mixed models. The algorithm is "automated" in the sense that at each iteration, the Monte Carlo error is assessed and the number of samples is increased if the change in parameter estimates from the previous iteration is "swamped" with Monte Carlo error. To make this possible, independently and identically distributed random samples are generated at each iteration, allowing one to use standard central limit theory to assess the Monte Carlo error. We now extend their algorithm to the multivariate generalized linear mixed models considered here.

The EM algorithm (Dempster et al. 1977) is an iterative method for obtaining maximum likelihood estimators in situations with incomplete data. For multivariate generalized linear mixed models, the random effects $\mathbf{u}' = (\mathbf{u}_1', \cdots, \mathbf{u}_n')$ are treated as the missing data. The EM algorithm is defined in terms of the complete log-likelihood

$$l_C(\Psi) = \log f\{\mathbf{y}, \mathbf{u}; \Psi\} = \sum_{i=1}^{n} \left[\log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) + \log g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)\right], \tag{3.39}$$

where $\Psi' = (\boldsymbol{\beta}', \mathrm{vech}(\Sigma)')$ and, for convenience, $f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) = \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i)$. The algorithm is divided into an Expectation Step (E-step) and a Maximization Step (M-step).

E-step

We begin by considering the E-step at the $(s+1)$th iteration. In the E-step, the expectation of the complete log-likelihood (3.39) is determined with respect to the conditional distribution $h(\mathbf{u} \mid \Psi^{(s)}; \mathbf{y})$. That is,

$$\begin{aligned} Q(\Psi \mid \Psi^{(s)}) &= E\{\log f(\mathbf{y}, \mathbf{u}; \Psi) \mid \Psi^{(s)}; \mathbf{y}\} \\ &= \sum_{i=1}^{n} \int \cdots \int \left[\log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) + \log g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)\right] \; h(\mathbf{u}_i \mid \Psi^{(s)}; \mathbf{y}_i) \, d\mathbf{u}_i \end{aligned} \tag{3.40}$$

The expectation in (3.40) can not be obtained in closed form since the conditional distribution $h(\mathbf{u}_i \mid \Psi^{(s)}; \mathbf{y}_i)$ contains the multivariate normal density. However, one can use Monte Carlo methods to approximate this expectation by generating samples from $h(\mathbf{u}_i \mid \Psi^{(s)}; \mathbf{y}_i)$. For the Monte Carlo EM algorithm proposed by Booth and Hobert (1999), the generated samples used to approximate (3.40) must be independently and identically distributed. They suggested either importance sampling or rejection samples to generate such samples, of which we utilize the latter approach for our algorithm.

The following multivariate rejection procedure (Geweke 1996) can be used to select K random samples $\boldsymbol{\varsigma}_{il}^{(s)}$, $l = 1, \cdots, K$, from $h(\mathbf{u}_i \mid \Psi^{(s)}; \mathbf{y}_i)$:

1. Sample $\varsigma_{il}^{(s)}$ from $g_{\texttt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma^{(s)})$ and independently sample w from the uniform(0,1) distribution.
2. Accept if $w \leq f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}^{(s)}; \varsigma_{il}^{(s)})/\tau$, otherwise go to 1,

where $\tau = \sup_{\mathbf{u}} \; f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}^{(s)}; \mathbf{u}_i)$. To calculate τ, note that $f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}^{(s)}; \mathbf{u}_i)$ can be considered a multivariate generalized linear model with unknown parameter vector $\mathbf{u}_i$. Thus one can find the maximum likelihood estimate of $\mathbf{u}_i$, say $\hat{\mathbf{u}}_i$, by fitting a multivariate generalized linear model that includes an offset of $Z_{ij}\boldsymbol{\beta}^{(s)}$. The value of τ is then given by $\tau = f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}^{(s)}; \hat{\mathbf{u}}_i)$.

Given n sets of K multivariate samples from $h(\mathbf{u}_i \mid \boldsymbol{\Psi}^{(s)}; \mathbf{y}_i)$, $i = 1, \cdots, n$, a K multivariate sample approximation to $Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)})$ is given by

$$Q_K(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)}) = \frac{1}{K}\sum_{i=1}^{n}\sum_{l=1}^{K}\left[\log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \varsigma_{il}^{(s)}) + \log g_{\texttt{MVN}}(\varsigma_{il}^{(s)}; \mathbf{0}, \Sigma)\right] \tag{3.41}$$

We will consider the assessment of the Monte Carlo error in (3.41) below.

M-step

The M-step at the $(s+1)$th iteration consists of maximizing the Monte Carlo approximation $Q_K(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)})$ in (3.41) with respect to $\boldsymbol{\Psi}$. Since the elements $\boldsymbol{\beta}$ and Σ of $\boldsymbol{\Psi}$ occur separately in the two terms of (3.41), the terms can be maximized individually. Consider the first term

$$\frac{1}{K}\sum_{i=1}^{n}\sum_{l=1}^{K}\log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \varsigma_{il}^{(s)}), \tag{3.42}$$

and note that (3.42) is a multivariate generalized linear model. By replicating the data vectors $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ K times, the linear predictor for the lth multivariate sample of the jth observation on the ith subject can be written

$$\boldsymbol{\eta}_{ijl} = Z_{ijl}\boldsymbol{\beta} + W_{ijl}\varsigma_{il}^{(s)}.$$

Maximization with respect to $\boldsymbol{\beta}$ can be carried out by Fisher scoring with $W_{ijl}\varsigma_{il}^{(s)}$ as an offset term. The same Fisher scoring algorithm given in Section 2.3 applies here, with an additional summation over l in (2.6) and (2.7).

The second term to be maximized in (3.41) is given by

$$\frac{1}{K}\sum_{i=1}^{n}\sum_{l=1}^{K}\log g_{\text{MVN}}(\varsigma_{il}^{(s)}; \mathbf{0}, \Sigma), \tag{3.43}$$

which is just the log of a multivariate normal density. For unstructured covariance matrices or when the random effects are assumed independent, closed form solutions for the maximum likelihood estimate of Σ exist. Generally, one can use an iterative procedure to find the maximum of (3.43) with respect to Σ (see, e.g., Searle et al. 1992, Chap. 11).

Monte Carlo error of $\boldsymbol{\Psi}^{(s+1)}$

The approximation of $Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)})$ in (3.41) inevitably will contain Monte Carlo error. This error is propagated through to the M-step, resulting in incorrect estimates for $\boldsymbol{\beta}$ and Σ. The Monte Carlo error can be made smaller by approximating $Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}^{(s)})$ with larger multivariate samples K. However, the larger K is chosen, the longer the estimation routine will take to converge. Booth and Hobert (1999) proposed a method for evaluating the Monte Carlo error in the estimates of $\boldsymbol{\beta}$ and Σ. Using the error estimate, they proposed a way to automate the choice of K at each iteration. Thus, they could start with a smaller K, when the parameter estimates were far from the maximum likelihood estimates, and increase K as they neared the maximum likelihood estimates. In addition, after taking into account the Monte Carlo error in the parameter estimates, they could accurately evaluate the convergence of the parameter estimates.

Following Booth and Hobert (1999), define $Q^{(j)}(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}') = \frac{d^j}{d\boldsymbol{\Psi}^j}\, Q(\boldsymbol{\Psi} \mid \boldsymbol{\Psi}')$ for $j = 1, 2$ and define $Q_K^{(j)}$ similarly. Booth and Hobert (1999) showed that conditional on $\boldsymbol{\Psi}^{(s)}$, $\boldsymbol{\Psi}^{(s+1)}$ is approximately normal with mean $\boldsymbol{\Psi}^{*(s+1)}$ and variance

$$\begin{aligned} &\operatorname{var}(\boldsymbol{\Psi}^{(s+1)} \mid \boldsymbol{\Psi}^{(s)}) \approx \\ &\qquad Q_K^{(2)}(\boldsymbol{\Psi}^{*(s+1)} \mid \boldsymbol{\Psi}^{(s)})^{-1} \operatorname{var}\left\{Q_K^{(1)}(\boldsymbol{\Psi}^{*(s+1)} \mid \boldsymbol{\Psi}^{(s)})\right\} Q_K^{(2)}(\boldsymbol{\Psi}^{*(s+1)} \mid \boldsymbol{\Psi}^{(s)})^{-1}, \end{aligned} \tag{3.44}$$

where $\boldsymbol{\Psi}^{*(s+1)}$ satisfies $Q_K^{(1)}(\boldsymbol{\Psi}^{*(s+1)} \mid \boldsymbol{\Psi}^{(s)}) = 0$. An estimate of (3.44) is obtained by substituting $\boldsymbol{\Psi}^{(s+1)}$ for $\boldsymbol{\Psi}^{*(s+1)}$, and estimating $\operatorname{var}\left\{Q_K^{(1)}(\boldsymbol{\Psi}^{*(s+1)} \mid \boldsymbol{\Psi}^{(s)})\right\}$ with

$$\begin{aligned} &\widehat{\operatorname{var}}\left\{Q_K^{(1)}(\boldsymbol{\Psi}^{*(s+1)} \mid \boldsymbol{\Psi}^{(s)})\right\} = \\ &\qquad \frac{1}{K^2} \sum_{l=1}^{K} \left\{\frac{d}{d\boldsymbol{\Psi}} \log f(\mathbf{y}, \boldsymbol{\varsigma}_l^{(s)}; \boldsymbol{\Psi}^{(s+1)})\right\} \left\{\frac{d}{d\boldsymbol{\Psi}} \log f(\mathbf{y}, \boldsymbol{\varsigma}_l^{(s)}; \boldsymbol{\Psi}^{(s+1)})\right\}', \end{aligned} \tag{3.45}$$

where $\boldsymbol{\varsigma}_l^{(s)'} = (\boldsymbol{\varsigma}_{1l}^{(s)'}, \cdots, \boldsymbol{\varsigma}_{nl}^{(s)'})$, $l = 1, \cdots, K$ are the K sets of n multivariate random samples for the sth iteration. For the multinomial random effects models, the elements of (3.44) and (3.45) have the forms given below.

We first consider $\frac{d}{d\boldsymbol{\Psi}} \log f(\mathbf{y}, \boldsymbol{\varsigma}_l^{(s)}; \boldsymbol{\Psi}^{(s+1)})$, which is the first derivative of the complete log-likelihood (3.39) with respect to $\boldsymbol{\Psi}$, evaluated at $\boldsymbol{\Psi}^{(s+1)}$ and the random sample $\boldsymbol{\varsigma}_l^{(s)}$. Only the first term in (3.39) contains $\boldsymbol{\beta}$, and it has the form of a multivariate generalized linear model. Thus the derivative of $\log f(\mathbf{y}, \boldsymbol{\varsigma}_l^{(s)}; \boldsymbol{\Psi}^{(s+1)})$ with respect to $\boldsymbol{\beta}$ is just the score function:

$$\frac{d}{d\boldsymbol{\beta}} \log f(\mathbf{y}, \boldsymbol{\varsigma}_l^{(s)}; \boldsymbol{\Psi}^{(s+1)}) = \sum_{i=1}^{n} \sum_{j=1}^{T_i} Z_{ij}' \, D_{ij} \, R_{\boldsymbol{\pi}_{ij}}^{-1} \, (\bar{\mathbf{y}}_{ij} - \boldsymbol{\pi}_{ij}), \tag{3.46}$$

where D_{ij}, $R_{\boldsymbol{\pi}_{ij}}$, and $\boldsymbol{\pi}_{ij}$ are calculated using $\boldsymbol{\beta}^{(s+1)}$ and $\boldsymbol{\varsigma}_l^{(s)}$. The second term in (3.39) is the log of a multivariate normal density, which has a derivative with respect to $\operatorname{vech}(\Sigma)$ of

$$\begin{aligned} \frac{d}{d\operatorname{vech}(\Sigma)} \log f(\mathbf{y}, \boldsymbol{\varsigma}_l^{(s)}; \boldsymbol{\Psi}^{(s+1)}) &= \sum_{i=1}^{n} \left[-\Sigma^{(s+1)^{-1}} + \frac{1}{2} \operatorname{diag}\left(\Sigma^{(s+1)^{-1}}\right)\right. \\ &\left. + \Sigma^{(s+1)^{-1}} \, \boldsymbol{\varsigma}_{il}^{(s)} \, \boldsymbol{\varsigma}_{il}^{(s)'} \, \Sigma^{(s+1)^{-1}} - \frac{1}{2} \operatorname{diag}\left(\Sigma^{(s+1)^{-1}} \, \boldsymbol{\varsigma}_{il}^{(s)} \, \boldsymbol{\varsigma}_{il}^{(s)'} \, \Sigma^{(s+1)^{-1}}\right)\right]. \end{aligned} \tag{3.47}$$

The derivatives (3.46) and (3.47) are stacked into a column vector and used in (3.45).

To estimate (3.44), we also need to calculate $Q_K^{(2)}(\boldsymbol{\Psi}^{(s+1)} \mid \boldsymbol{\Psi}^{(s)})$ which corresponds to the second derivative matrix of (3.41) with respect to $\boldsymbol{\Psi}\ \boldsymbol{\Psi}'$ and has the form

$$\frac{d^2}{d\boldsymbol{\Psi}\, d\boldsymbol{\Psi}'}\, Q_K(\boldsymbol{\Psi}^{(s+1)} \mid \boldsymbol{\Psi}^{(s)}) = \begin{bmatrix} \frac{d^2}{d\boldsymbol{\beta}\, d\boldsymbol{\beta}'}\, Q_K & 0 \\ 0 & \frac{d^2}{d\,\mathrm{vech}(\Sigma)\, d\,\mathrm{vech}(\Sigma)'}\, Q_K \end{bmatrix}. \tag{3.48}$$

Note that the off-diagonal elements are zero since $\boldsymbol{\beta}$ and Σ do not occur together in either the first or second term of (3.41). Replicating $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ K times and exploiting that fact that the first term is a multivariate generalized linear model, the second derivative of $Q_K(\boldsymbol{\Psi}^{(s+1)} \mid \boldsymbol{\Psi}^{(s)})$ with respect to $\boldsymbol{\beta}$ is just the observed information matrix,

$$\frac{d^2}{d\boldsymbol{\beta}\, d\boldsymbol{\beta}'}\, Q_K = \frac{1}{K} \sum_{i=1}^{n} \sum_{l=1}^{K} \sum_{j=1}^{T_i} \Bigg\{ Z'_{ijl}\, D_{ijl}\, R^{-1}_{\boldsymbol{\pi}_{ijl}}\, D'_{ijl}\, Z_{ijl} \\ - \left[\sum_{r=1}^{q} Z'_{ijl} U_{ijlr} Z_{ijl} (\bar{y}_{ijlr} - \pi_{ijlr}) \right] \Bigg\}, \tag{3.49}$$

where D_{ijl}, $R_{\boldsymbol{\pi}_{ijl}}$, U_{ijlr}, and π_{ijlr} are computed using $\boldsymbol{\Psi}^{(s+1)}$ and $\varsigma_l^{(s)}$, and U_{ijlr} is defined as in Chapter 2.

The second element in (3.48) is the second derivative of the log of a multivariate normal density with respect to the unique elements of Σ. To provide an explicit formula for calculating this derivative matrix, we first need the following notation.

Notation. (Graham 1981)

E_{rt} : A q by q matrix of zeros with a one in the (r, t)th cell.

I_q : A q by q identity matrix.

J_q : A q by q matrix of ones.

δ_{rt} $= 1$ if $r \neq t$, 0 otherwise.

U $= \sum_{r=1}^{q} \sum_{t=1}^{q} E_{rt} \otimes E_{st}$.

$\bar{U}$ $= \sum_{r=1}^{q} \sum_{t=1}^{q} E_{rt} \otimes E_{rt}$.

$$\bar{U}^{\star} \;\; = \bar{U} + U - \textstyle\sum_{r=1}^{q} E_{rr} \otimes E_{rr}.$$

We also denote element-wise multiplication between two conformable matrices by the symbol $\cdot *$. Define the matrix A_{il} by

$$\begin{aligned} A_{il} = \sum_{r=1}^{q}\sum_{t=1}^{q}\Bigg\{ & \left[E_{rt} \otimes \left(\Sigma^{(s+1)^{-1}} E_{rt}\right)\right] \otimes \left[I_q \otimes \left(\Sigma^{(s+1)^{-1}} \varsigma_{il}^{(s)} \varsigma_{il}^{(s)'} \Sigma^{(s+1)^{-1}}\right)\right] \\ & + \delta_{rt}\left[E_{rt} \otimes \left(\Sigma^{(s+1)^{-1}} E_{tr}\right)\right]\left[I_q \otimes \left(\Sigma^{(s+1)^{-1}} \varsigma_{il}^{(s)} \varsigma_{il}^{(s)'} \Sigma^{(s+1)^{-1}}\right)\right] \\ & + \left[I_q \otimes \left(\Sigma^{(s+1)^{-1}} \varsigma_{il}^{(s)} \varsigma_{il}^{(s)'} \Sigma^{(s+1)^{-1}}\right)\right]\left[E_{rt} \otimes \left(E_{rt}\Sigma^{(s+1)^{-1}}\right)\right] \\ & + \delta_{rt}\left[I_q \otimes \left(\Sigma^{(s+1)^{-1}} \varsigma_{il}^{(s)} \varsigma_{il}^{(s)'} \Sigma^{(s+1)^{-1}}\right)\right]\left[E_{rt} \otimes \left(E_{tr}\Sigma^{(s+1)^{-1}}\right)\right]\Bigg\}. \end{aligned} \quad (3.50)$$

Then

$$\begin{aligned} \frac{d^2}{d\,\mathrm{vech}(\Sigma)\, d\,\mathrm{vech}(\Sigma)'} Q_K = \frac{1}{K}\sum_{i=1}^{n}\sum_{l=1}^{K}\Bigg\{ & \left[I_q \otimes \Sigma^{(s+1)^{-1}}\right] \bar{U}^{\star} \left[I_q \otimes \Sigma^{(s+1)^{-1}}\right] \\ - \frac{1}{2}\left[J_q \otimes I_q\right] \cdot * \left[I_q \otimes \Sigma^{(s+1)^{-1}}\right] \bar{U}^{\star} \left[I_q \otimes \Sigma^{(s+1)^{-1}}\right] & - A_{il} + \frac{1}{2}\left[J_q \otimes I_q\right] \cdot * A_{il}\Bigg\}. \end{aligned} \quad (3.51)$$

Using (3.49) and (3.51) in (3.48), one obtains an estimate for $Q_K^{(2)}(\Psi^{(s+1)} \mid \Psi^{(s)})$.

Algorithm automation

The estimate of the variance of $\Psi^{(s+1)}$ given $\Psi^{(s)}$ can now be used to construct a rule for updating the Monte Carlo sample size K. Consider the difference $||\Psi^{*(s+1)} - \Psi^{(s)}||$, which is the difference between the true maximizer of $Q^{(1)}$ at the $(s+1)$th iteration and the estimated value at the (s)th iteration. If this difference is small relative to the Monte Carlo error associated with $\Psi^{(s+1)}$, then $\Psi^{(s+1)}$ will be of no use for estimating $\Psi^{*(s+1)}$ as it will be overwhelmed with error. At such a point, one would need to take a larger Monte Carlo sample K to help reduce the error. Booth and Hobert (1999) suggested calculating a confidence ellipsoid about $\Psi^{*(s+1)}$ and checking if the previous estimate $\Psi^{(s)}$ was contained in the region. If $\Psi^{(s)}$ was contained in the ellipsoid, they concluded that the current estimate $\Psi^{(s+1)}$ was swamped with Monte

Carlo error and that K should be increased. Specifically,

$$\text{if} \quad (\boldsymbol{\Psi}^{(s+1)} - \boldsymbol{\Psi}^{(s)})' \, \widehat{\text{var}}(\boldsymbol{\Psi}^{(s+1)} \mid \boldsymbol{\Psi}^{(s)})^{-1} \, (\boldsymbol{\Psi}^{(s+1)} - \boldsymbol{\Psi}^{(s)}) \leq \chi^2_{\text{df},1-\alpha},$$
$$\text{then} \quad K = K + \frac{K}{\Delta}, \tag{3.52}$$

where df denotes the dimension of $\boldsymbol{\Psi}^{(s+1)}$. The most appropriate choices for Δ and α are still under research, however, we have had success with $\Delta = 4$ and $\alpha = 0.25$.

To determine convergence, we utilized a standard stopping rule. Denote the dimension of $\boldsymbol{\Psi}$ by p^*. Then the convergence criterion is satisfied if

$$\max_{i=1,\cdots,p^*} \frac{|\Psi_i^{(s+1)} - \Psi_i^{(s)}|}{|\Psi_i^{(s)}| + 0.001} < 0.002. \tag{3.53}$$

As the EM algorithm can be extremely slow to converge when a parameter is near the boundary of the parameter space (e.g., when a variance component is near zero), one should also monitor the parameter estimates during the algorithm to ensure this is not occurring. We programmed the Monte Carlo EM algorithm using the Ox programming language.

3.4 Inference and Prediction

We now consider inference for the fixed effects $\boldsymbol{\beta}$ in the multivariate generalized linear mixed model, as well as prediction of the random effects $\mathbf{u}_i$, $i = 1, \cdots, n$. As the estimates obtained from the quasi-Newton algorithm and Monte Carlo EM algorithm are approximate maximum likelihood, inference concerning the fixed effects is based on the usual asymptotic maximum likelihood theory. Before briefly discussing these approaches, such as the Wald test and the likelihood-ratio test, we first outline the calculation of standard error estimates for the algorithms considered in the previous section. We conclude this section by showing how adaptive Gauss-Hermite quadrature can be used to obtain predictions for the random effects.

3.4.1 Standard Errors

We now provide the necessary formulas for obtaining standard errors of $\hat{\boldsymbol{\Psi}}$ upon convergence of the algorithms in the previous section. We obtain these estimates by inverting the observed information matrix, which requires calculation of the second derivative of marginal log-likelihood of the general multinomial random effects model (i.e. the log of (3.19)) with respect to $\boldsymbol{\Psi}\,\boldsymbol{\Psi}'$. Hedeker and Gibbons (1994) based standard errors on the expected information matrix for their general ordinal random effects model. For random effects models, the observed information matrix is often easier to calculate. Efron and Hinkley (1978) argued that standard errors based on the observed information matrix are also "closer" to the real data. However, the observed information matrix is not guaranteed to be positive definite, as is the case for the expected information matrix. Tutz and Hennevogl (1996) based standard errors on the estimated information matrix $\sum_{i=1}^{n} s_i(\hat{\boldsymbol{\Psi}})s_i(\hat{\boldsymbol{\Psi}})'$ where $s_i(\hat{\boldsymbol{\Psi}})$ is the contribution of the ith cluster to the approximated score function. They warned, however, that this approach has a tendency of overestimating the true standard errors. In fact, we will see in Section 3.6.1 that this approach can be extremely inaccurate. Denoting the marginal log-likelihood by $l(\boldsymbol{\Psi})$, the matrix of second derivatives has the form

$$\frac{d^2 l(\boldsymbol{\Psi})}{d\boldsymbol{\Psi}\, d\boldsymbol{\Psi}'} = \begin{bmatrix} \frac{d^2 l(\boldsymbol{\Psi})}{d\boldsymbol{\beta}\, d\boldsymbol{\beta}'} & \frac{d^2 l(\boldsymbol{\Psi})}{d\boldsymbol{\beta}\, d\,\mathrm{vech}(\Sigma)'} \\ \frac{d^2 l(\boldsymbol{\Psi})}{d\,\mathrm{vech}(\Sigma)\, d\boldsymbol{\beta}'} & \frac{d^2 l(\boldsymbol{\Psi})}{d\,\mathrm{vech}(\Sigma)\, d\,\mathrm{vech}(\Sigma)'} \end{bmatrix}. \tag{3.54}$$

As each component of (3.54) contains intractable integrals, adaptive Gauss-Hermite quadrature or Monte Carlo integration is needed. For the EM algorithm, Louis (1982) showed how the observed information matrix can be obtained using items already calculated in the EM-steps. Thus, for the Monte Carlo EM algorithm we use Louis' approach for obtaining standard errors, while for the quasi-Newton algorithm we directly compute (3.54).

We first consider direct approximation of (3.54) using adaptive Gauss-Hermite quadrature. Denote the marginal log-likelihood $l(\mathbf{\Psi})$ as

$$l(\mathbf{\Psi}) = \sum_{i}^{n} \log L_i, \tag{3.55}$$

where L_i is given in (3.25). Interchanging integrals and derivatives, the elements of (3.54) can be written in the form

$$\sum_{i=1}^{n} \frac{L\, B_i^{(e,f)} - C_i^{(e,f)}\, D_i^{(e,f)}}{L^2}, \quad e, f = 1, 2, \tag{3.56}$$

where $L = \sum_{i=1}^{n} L_i$. For the (1,1) element, that is $\dfrac{d^2 l(\mathbf{\Psi})}{d\boldsymbol{\beta}\, d\boldsymbol{\beta}'}$,

$$\begin{aligned} B_i^{(1,1)} = & \int \cdots \int f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma) \\ & \times \left[\frac{d^2 \log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i)}{d\boldsymbol{\beta}\, d\boldsymbol{\beta}'} + \frac{d \log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i)}{d\boldsymbol{\beta}} \frac{d \log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i)}{d\boldsymbol{\beta}'} \right] d\mathbf{u}_i, \end{aligned} \tag{3.57}$$

$$C_i^{(1,1)} = \int \cdots \int f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma) \frac{d \log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i)}{d\boldsymbol{\beta}} d\mathbf{u}_i, \tag{3.58}$$

and $D_i^{(1,1)} = C_i^{(1,1)'}$.

For the (2,2) element, that is $\dfrac{d^2 l(\mathbf{\Psi})}{d\mathrm{vech}(\Sigma)\, d\mathrm{vech}(\Sigma)'}$,

$$\begin{aligned} B_i^{(2,2)} = & \int \cdots \int f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma) \\ & \times \left[\frac{d^2 \log g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)}{d\mathrm{vech}(\Sigma)\, d\mathrm{vech}(\Sigma)'} + \frac{d \log g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)}{d\mathrm{vech}(\Sigma)} \frac{d \log g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)}{d\mathrm{vech}(\Sigma)'} \right] d\mathbf{u}_i, \end{aligned} \tag{3.59}$$

$$C_i^{(2,2)} = \int \cdots \int f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma) \frac{d \log g_{\mathtt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)}{d\mathrm{vech}(\Sigma)} d\mathbf{u}_i, \tag{3.60}$$

and $D_i^{(2,2)} = C_i^{(2,2)'}$.

Lastly, for the (1,2) element, that is $\dfrac{d^2 l(\boldsymbol{\Psi})}{d\boldsymbol{\beta}\, d\text{vech}(\Sigma)'}$,

$$B_i^{(1,2)} = \int \cdots \int f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i) g_{\texttt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma) \frac{d \log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}; \mathbf{u}_i)}{d\boldsymbol{\beta}} \frac{d \log g_{\texttt{MVN}}(\mathbf{u}_i; \mathbf{0}, \Sigma)}{d\text{vech}(\Sigma)'} d\mathbf{u}_i, \tag{3.61}$$

$C_i^{(1,2)} = C_i^{(1,1)}$, and $D_i^{(1,2)} = D_i^{(2,2)'}$.

Thus for each element in (3.54) there are two sets of integrals to approximate. The adaptive Gauss-Hermite approximation of the first set of integrals found in the marginal likelihood L has already been given in (3.29) and is calculated here using the final parameter estimates $\hat{\boldsymbol{\Psi}}$. Let $\boldsymbol{\varsigma}_{\mathbf{l}}^{\star'} = (\varsigma_{1\mathbf{l}}^{\star'}, \cdots, \varsigma_{n\mathbf{l}}^{\star'})$, $l = 1, \cdots, K$, denote the centered and scaled nodes used in approximating L, with corresponding curvatures $(\Sigma_1^*, \cdots, \Sigma_n^*)$. We then approximate the second set of integrals by using these same nodes $\boldsymbol{\varsigma}_{\mathbf{l}}^{\star}$ along with the final parameter estimates. This parallels what is done in Monte Carlo methods. That is, random samples drawn from the marginal log-likelihood at the final iteration are used to evaluate the observed information matrix. We have attempted to approximate each set of integrals individually. However the integrand of, for example, (3.57) is often ill-behaved causing adaptive Gauss-Hermite quadrature to perform poorly. We have already given approximation formulas for some of the remaining terms in (3.56). The approximations for $\{C_i^{(1,1)}, C_i^{(1,2)}, D_i^{(1,1)}\}$ and $\{C_i^{(2,2)}, D_i^{(1,2)}, D_i^{(2,2)}\}$ can be obtained from slight modifications of (3.35) and (3.38), respectively. Replicating the data vectors $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ K^m times, the adaptive quadrature approximation of $B_i^{(1,1)}$ is given by

$$B_i^{(1,1)} \approx \sum_{\mathbf{l}}^{K} c_{i\mathbf{l}}^{\star} \left\{ \left[\sum_{j=1}^{T_i} Z_{ij\mathbf{l}}' \, D_{ij\mathbf{l}} \, R_{\boldsymbol{\pi}_{ij\mathbf{l}}}^{-1} \, D_{ij\mathbf{l}}' \, Z_{ij\mathbf{l}} - \left(\sum_{r=1}^{q} Z_{ij\mathbf{l}}' U_{ij\mathbf{l}r} Z_{ij\mathbf{l}} (\bar{y}_{ij\mathbf{l}r} - \pi_{ij\mathbf{l}r}) \right) \right] \right.$$
$$\left. + \left[\sum_{\mathbf{l}}^{K} \sum_{j=1}^{T_1} Z_{ij\mathbf{l}}' \, D_{ij\mathbf{l}} \, R_{\boldsymbol{\pi}_{ij\mathbf{l}}}^{-1} \, (\bar{\mathbf{y}}_{ij\mathbf{l}} - \boldsymbol{\pi}_{ij\mathbf{l}}) \right] \left[\sum_{\mathbf{l}}^{K} \sum_{j=1}^{T_1} Z_{ij\mathbf{l}}' \, D_{ij\mathbf{l}} \, R_{\boldsymbol{\pi}_{ij\mathbf{l}}}^{-1} \, (\bar{\mathbf{y}}_{ij\mathbf{l}} - \boldsymbol{\pi}_{ij\mathbf{l}}) \right]' \right\}, \tag{3.62}$$

where summation is over $\mathbf{l} = (\mathbf{l}_1, \cdots, \mathbf{l}_m)$ and

$$c_{i\mathbf{l}}^{\star} = \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij\mathbf{l}} \mid \boldsymbol{\beta}; \varsigma_{i\mathbf{l}}^{\star}) \right] g_{\mathrm{MVN}}(\varsigma_{i\mathbf{l}}^{\star}; \mathbf{0}, \Sigma) \ \mid \Sigma_i^{*} \mid^{1/2} \ 2^{m/2} \ \varpi_{\mathbf{l}} \ \exp(\varsigma_{\mathbf{l}}' \, \varsigma_{\mathbf{l}}). \tag{3.63}$$

Also, the approximation of $B_i^{(1,2)}$ is given by

$$B_i^{(1,2)} \approx \sum_{\mathbf{l}}^{K} c_{i\mathbf{l}}^{\star} \left[\sum_{\mathbf{l}}^{K} \sum_{j=1}^{T_i} Z_{ij\mathbf{l}}' \, D_{ij\mathbf{l}} \, R_{\boldsymbol{\pi}_{ij\mathbf{l}}}^{-1} \, (\bar{\mathbf{y}}_{ij\mathbf{l}} - \boldsymbol{\pi}_{ij\mathbf{l}}) \right]$$
$$\times \left[-\Sigma^{-1} + \frac{1}{2} \operatorname{diag}(\Sigma^{-1}) + \Sigma^{-1} \, \varsigma_{i\mathbf{l}}^{\star} \, \varsigma_{i\mathbf{l}}^{\star\prime} \, \Sigma^{-1} - \frac{1}{2} \operatorname{diag}(\Sigma^{-1} \, \varsigma_{i\mathbf{l}}^{\star} \, \varsigma_{i\mathbf{l}}^{\star\prime} \, \Sigma^{-1}) \right]', \tag{3.64}$$

with $c_{i\mathbf{l}}^{\star}$ defined as in (3.63). The final approximation is that of $B_i^{(2,2)}$, which requires the second derivative of the log of a multivariate normal density with respect to the unique elements of Σ. We have already seen the general form of this derivative in (3.51), approximated by Monte Carlo methods. The form that is needed here is slightly different than (3.51), however. Let $M_{i\mathbf{l}}$ denote equation (3.51) with the $\frac{1}{K}$ and $\sum_{i=1}^{n}$ removed, summation over l changed to $\mathbf{l}$, and $\varsigma_{il}^{(s)}$ replaced with $\varsigma_{i\mathbf{l}}^{\star}$. Then the approximation of $B_i^{(2,2)}$ is given by

$$B_i^{(2,2)} \approx$$
$$\sum_{\mathbf{l}}^{K} c_{i\mathbf{l}}^{\star} \left\{ M_{i\mathbf{l}} + \left[-\Sigma^{-1} + \frac{1}{2} \operatorname{diag}(\Sigma^{-1}) + \Sigma^{-1} \, \varsigma_{i\mathbf{l}}^{\star} \, \varsigma_{i\mathbf{l}}^{\star\prime} \, \Sigma^{-1} - \frac{1}{2} \operatorname{diag}(\Sigma^{-1} \, \varsigma_{i\mathbf{l}}^{\star} \, \varsigma_{i\mathbf{l}}^{\star\prime} \, \Sigma^{-1}) \right] \right.$$
$$\left. \left[-\Sigma^{-1} + \frac{1}{2} \operatorname{diag}(\Sigma^{-1}) + \Sigma^{-1} \, \varsigma_{i\mathbf{l}}^{\star} \, \varsigma_{i\mathbf{l}}^{\star\prime} \, \Sigma^{-1} - \frac{1}{2} \operatorname{diag}(\Sigma^{-1} \, \varsigma_{i\mathbf{l}}^{\star} \, \varsigma_{i\mathbf{l}}^{\star\prime} \, \Sigma^{-1}) \right]' . \right\} \tag{3.65}$$

Using estimates (3.62), (3.64), and (3.65) along with the approximations for the other elements of (3.56), one can obtain an estimate of the observed information matrix (3.54) upon convergence. By inverting the negative of (3.54), the estimated asymptotic variance-covariance matrix for $\hat{\mathbf{\Psi}}$ is obtained.

In the context of the EM algorithm, Louis (1982) showed that the observed information matrix could be calculated from

$$\frac{d^2 l(\mathbf{\Psi})}{d\mathbf{\Psi}\, d\mathbf{\Psi}'} = Q^{(2)}(\mathbf{\Psi} \mid \hat{\mathbf{\Psi}}) + \mathrm{var}\{\log f(\mathbf{y}, \mathbf{u}; \mathbf{\Psi}) \mid \mathbf{y}; \hat{\mathbf{\Psi}}\}, \tag{3.66}$$

where the variance is with respect to $h(\mathbf{u} \mid \mathbf{\Psi}; \mathbf{y})$. Upon convergence, the final Monte Carlo sample, $\varsigma_l^\star$, along with the final parameter estimates, $\hat{\mathbf{\Psi}}$, can be used to estimate (3.66). Note that the Monte Carlo estimate of the first term in (3.66) was previously given in (3.48). The Monte Carlo approximation to the second term has the form

$$\begin{aligned} \mathrm{var}\{\log f(\mathbf{y}, \mathbf{u}; \mathbf{\Psi}) \mid \mathbf{y}; \hat{\mathbf{\Psi}}\} \approx {} & \frac{1}{K} \sum_{l=1}^{K} \left[\frac{d \log f(\mathbf{y}, \varsigma_l^\star; \hat{\mathbf{\Psi}})}{d\mathbf{\Psi}} \right] \left[\frac{d \log f(\mathbf{y}, \varsigma_l^\star; \hat{\mathbf{\Psi}})}{d\mathbf{\Psi}'} \right] \\ & - \left[\frac{1}{K} \sum_{l=1}^{K} \frac{d \log f(\mathbf{y}, \varsigma_l^\star; \hat{\mathbf{\Psi}})}{d\mathbf{\Psi}} \right] \left[\frac{1}{K} \sum_{l=1}^{K} \frac{d \log f(\mathbf{y}, \varsigma_l^\star; \hat{\mathbf{\Psi}})}{d\mathbf{\Psi}'} \right]. \end{aligned} \tag{3.67}$$

The necessary equations for the derivatives in (3.67) are given in (3.46) and (3.47). Taking the negative of (3.66) and inverting gives the desired estimate of the variance-covariance matrix.

3.4.2 Maximum Likelihood Inference

Inference concerning the fixed effects parameters in the multinomial random effects model is accomplished using standard asymptotic maximum likelihood theory. Such an approach is justified as long as the approximation based algorithms used are accurately approximating the intractable integrals. In theory one can choose a large enough Monte Carlo sample size or enough quadrature points to obtain the true maximum likelihood estimates. We assume that such accuracy has been reached. A thorough review of asymptotic maximum likelihood theory can be found in Prakasa Rao

(1987). We briefly review the theory for independent but not identically distributed random variables.

Let $\mathbf{y}' = (\mathbf{y}_1', \cdots, \mathbf{y}_n')$ be independent but not identically distributed random variables. Let $l_n(\boldsymbol{\beta})$, $s_n(\boldsymbol{\beta})$, $F_{E,n}(\boldsymbol{\beta})$, and $F_{O,n}(\boldsymbol{\beta})$ be the log-likelihood, score function, and expected and observed information matrices for the entire sample $\mathbf{y}$. Note that these equations are given in Section 2.2. We consider local maximum likelihood estimates $\hat{\boldsymbol{\beta}}_n$ for $\boldsymbol{\beta}$ in the interior of the parameter space. Though there is no guarantee that a maximum of $l_n(\boldsymbol{\beta})$ will exist or that local and global maxima will coincide, for many important models local and global maxima are identical and unique if they exist (see, e.g., Kaufmann 1988 for the consideration of multicategorical models). The standard $n^{1/2}$-asymptotics that we discuss hold under typical regularity conditions, one such condition being that $F_{E,n}(\boldsymbol{\beta})/n$ converges to a positive definite limit:

$$F_{E,n}(\boldsymbol{\beta})/n = E[F_{O,n}(\boldsymbol{\beta})]/n \rightarrow F(\boldsymbol{\beta}).$$

The following asymptotic results can be shown to hold under the regularity assumptions. The score function $s_n(\boldsymbol{\beta})$ is asymptotically normal

$$n^{-1/2} s_n(\boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, F(\boldsymbol{\beta})). \tag{3.68}$$

The maximum likelihood estimate $\hat{\boldsymbol{\beta}}_n$ asymptotically exists and is asymptotically consistent and normal

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N(\mathbf{0}, F(\boldsymbol{\beta})^{-1}). \tag{3.69}$$

Using these results, one can obtain asymptotic distributions for the likelihood-ratio, Wald, and score statistics.

Consider testing the following linear hypothesis:

$$H_o : C\boldsymbol{\beta} = \boldsymbol{\beta}_o \quad \text{versus} \quad H_a : C\boldsymbol{\beta} \neq \boldsymbol{\beta}_o,$$

where C has full row rank $s \leq p$, the dimension of $\boldsymbol{\beta}$. The likelihood-ratio statistic,

$$\lambda_{lrt} = -2[l_n(\hat{\boldsymbol{\beta}}_n) - l_n(\tilde{\boldsymbol{\beta}}_n)], \tag{3.70}$$

compares the likelihood value under the alternative hypothesis where $\tilde{\boldsymbol{\beta}}_n$ is the maximum likelihood estimate, to the likelihood value under the null hypothesis where $\hat{\boldsymbol{\beta}}_n$ is the maximum likelihood estimate. The Wald statistic,

$$\lambda_W = (C\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_o)' \; [C \; F_E^{-1}(\hat{\boldsymbol{\beta}}_n) \; C']^{-1} \; (C\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_o), \tag{3.71}$$

compares the distance between the unrestricted estimate $C\hat{\boldsymbol{\beta}}_n$ and its value under the null hypothesis. The score statistic,

$$\lambda_S = s'(\tilde{\boldsymbol{\beta}}_n) \; F^{-1}(\tilde{\boldsymbol{\beta}}_n) \; s(\tilde{\boldsymbol{\beta}}_n), \tag{3.72}$$

compares the score function for the unrestricted model evaluated under the null hypothesis maximum likelihood estimate $\tilde{\boldsymbol{\beta}}_n$ to zero. Asymptotic χ^2 distributions of the likelihood-ratio, Wald, and score statistics can be derived by expanding the log-likelihood $l_n(\boldsymbol{\beta})$ in a Taylor series about $\hat{\boldsymbol{\beta}}_n$ and using results (3.68) and (3.69). Asymptotically the three tests are equivalent under the null and have the same limiting χ_s^2 distribution for the hypothesis given above.

For the multinomial random effects models, the observed information matrix is easier to calculate then the expected information matrix. Since $F_{E,n}(\boldsymbol{\beta})/n$ and $E[F_{O,n}(\boldsymbol{\beta})]/n$ converge to the same positive definite limit, the observed information matrix can be inserted into (3.71) and (3.72). Only approximations are available for the components in (3.70), (3.71), and (3.72), and we assume that they have been adequately approximated. For testing of variance components, the asymptotics for the Wald test and likelihood-ratio test break down (Self and Liang 1987; Bryk and Raudenbush 1992, p. 55), as the test involves the boundary of the parameter space. The score test, however, is not affected by such conditions (Chant 1974). In Chapter 5 we will consider two approximate score tests for testing individual variance components.

3.4.3 Prediction

Besides estimation of the fixed effects parameter vector $\boldsymbol{\beta}$, one might also be interested in predicting the values of the random effects vector $\mathbf{u}_i$, $i = 1, \cdots, n$, or linear combinations of fixed and random effects. Such predictions are based on the conditional expectation of the random effects given the data and the final parameter estimates. For the multinomial random effects model, this expectation is of the form

$$E[\mathbf{u}_i \mid \mathbf{y}_i, \hat{\boldsymbol{\Psi}}] = \frac{\int \cdots \int \mathbf{u}_i f(\tilde{\mathbf{y}}_i \mid \hat{\boldsymbol{\beta}}; \mathbf{u}_i)\; g_{\text{MVN}}(\mathbf{u}_i; \mathbf{0}, \hat{\Sigma})\; d\mathbf{u}_i}{\int \cdots \int f(\tilde{\mathbf{y}}_i \mid \hat{\boldsymbol{\beta}}; \mathbf{u}_i)\; g_{\text{MVN}}(\mathbf{u}_i; \mathbf{0}, \hat{\Sigma})\; d\mathbf{u}_i}, \tag{3.73}$$

where $\mathbf{y}_i' = (\mathbf{y}_{i1}', \cdots, \mathbf{y}_{iT_i}')$. The expectation in (3.73) requires integral approximations, thus adaptive Gauss-Hermite quadrature or Monte Carlo integration can be employed. Though (3.73) involves only the data for subject i, the estimates of the $\mathbf{u}_i$ "borrow" information from all of the subjects since $\hat{\boldsymbol{\Psi}}$ is obtained from the complete data.

Booth and Hobert (1998) discussed the calculation of standard errors for predictions involving random effects. For linear mixed models, standard errors of prediction are typically based on $\text{Var}(\mathbf{u}_i \mid \hat{\boldsymbol{\Psi}}; \mathbf{y}_i)$. Booth and Hobert (1998) showed that this approach was inappropriate for mixed models with non-normal responses, and proposed the use of the conditional mean squared error of prediction (CMSEP). The CMSEP takes into account the variability associated with $\hat{\boldsymbol{\Psi}}$. A Taylor series approximation to this correction factor can be found in Booth and Hobert (1998).

3.5 Pseudo-Likelihood Estimation

The estimation methods proposed in Section 3.3 utilized numerical integration techniques to carry out maximum likelihood estimation. Though based on approximations, estimates obtained from these methods can be considered "exact" maximum likelihood, since, in theory, one can increase the Monte Carlo sample size or number of quadrature points until a desired accuracy is reached. In contrast, we now consider an approximate method for finding maximum likelihood estimates for the multinomial random effects model. An advantage of the approximate method is that it avoids the intractable integrals completely. Thus, estimation does not require the computationally intensive, numerical integration methods needed in Section 3.3. The estimation routine is also attractive, as it patterns that used in standard linear mixed models. We begin by reviewing some of the recent literature on approximate methods for generalized linear mixed models. We then present the estimation routine for the multinomial random effects model. We conclude this section with some discussion concerning the proposed model.

3.5.1 Approximate Inference in Generalized Linear Mixed Models

There have been a number of proposals for approximate maximum likelihood estimation in generalized linear mixed models. Breslow and Clayton (1993) proposed a penalized quasi-likelihood (PQL) approach for fitting generalized linear mixed models. Under the assumption of normality for the random effects, they replaced the integrated quasi-likelihood with a quadratic Laplace approximation in terms of the current estimates of the random effects. Certain terms in the approximation were assumed to vary slowly enough, as a function of the mean of the generalized linear model, so that they could be ignored. The quasi-likelihood deviance terms in the approximation were also replaced with estimated Pearson residuals. Breslow and Clayton (1993) proposed a Fisher scoring algorithm for estimation of the fixed and random effects and REML estimation for the variance component estimation. The estimating equations corresponded to those obtained by Harville (1977) for best linear unbiased estimation in the associated normal theory model. In addition, Breslow and Clayton (1993) proposed a marginal quasi-likelihood (MQL) approach for modeling the marginal mean in a generalized linear mixed model. The PQL and MQL approaches correspond to the subject-specific and population-averaged approaches of Zeger et al. (1988).

Engel and Keen (1994) proposed a similar method to that of the PQL approach. Motivated from a quasi-likelihood, they utilized iteratively re-weighted least squares and iterated MINQUE to estimate the fixed and random effects, and the variance components. In contrast to Breslow and Clayton (1993), Engel and Keen (1994) allowed for an additional overdispersion parameter. They used a method of moments estimator for updating the overdispersion parameter in which they equated the Pearson's chi-square statistic to its degrees of freedom.

Wolfinger and O'Connell (1993) presented an approximate method for fitting generalized linear mixed models using a pseudo-likelihood (PL) approach. A PL procedure is based on the following concept. For parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, PL estimates of $\boldsymbol{\theta}$ are found by treating the parameters $\boldsymbol{\beta}$ as known and equal to their current value, and then estimating $\boldsymbol{\theta}$ by maximum likelihood. Wolfinger and O'Connell (1993) considered a model that allowed covariance structures for modeling both population-averaged and subject-specific associations. In terms of the PL concept, the elements of the covariance matrices corresponded to $\boldsymbol{\theta}$ and the fixed and random effects corresponded to $\boldsymbol{\beta}$. In their approach, the fixed and random effects were estimated from a linear mixed model based on an approximately normal pseudo-response variable. Then the elements of the covariance matrices were updated using either maximum likelihood or REML. Similarly to Engel and Keen (1994), they also allowed for an additional overdispersion parameter. When the overdispersion parameter is forced to be 1.0 and the population-averaged covariance matrix is ignored, the PQL approach of Breslow and Clayton (1993) and Engel and Keen (1994), and the PL approach of Wolfinger and O'Connell (1993) are equivalent.

Keen and Engel (1997) extended the methods used in Engel and Keen (1994) to threshold models for ordinal responses. The threshold models were motivated from an underlying continuous response which followed a linear mixed model (see Section 3.2.2). They assumed that the residuals for the underlying linear mixed model were normally distributed, resulting in a cumulative probit random effects model. Additionally, they considered a more general class of link functions in which the distribution of the residuals was allowed to depend on a set of additional shape parameters. Specifically, they assumed that the residuals followed a t-distribution. Keen and Engel (1997) showed that improved fits for particular datasets could be just as easily obtained by changing the link function for the cumulative probabilities as by introducing variability in the thresholds.

Though attractive for their computational simplicity, these approximate methods for generalized linear mixed models have been shown to be biased for Bernoulli and binomial response data. Breslow and Clayton (1993) reported that the accuracy of the regression coefficients improved as the binomial denominators increased. Breslow and Lin (1995) and Lin and Breslow (1996) studied the asymptotic bias of the variance components and regression coefficients, and proposed bias correction factors for adjusting these estimates. Engel (1998) used a simple probit-normal model with two Bernoulli observations per cluster and the overall mean as the only fixed effect to show the severity in bias that can occur in the variance component estimation. In general, the approximate methods such as PQL will perform adequately when the binomial sample sizes are large and the variance components are small to moderate in size (Engel and Keen 1994).

3.5.2 Pseudo-Likelihood Estimation for Multinomial Random Effects Models

We now generalize the PL estimation approach of Wolfinger and O'Connell (1993) to multivariate generalized linear mixed models for nominal and ordinal response data. To motivate the PL algorithm, we consider the model for the complete response vector

$\bar{\mathbf{y}} = [\bar{\mathbf{y}}_{ij}]$,

$$\bar{\mathbf{y}} = \boldsymbol{\pi} + \mathbf{e}, \tag{3.74}$$

with link function given by

$$\mathbf{g}(\boldsymbol{\pi}) = Z\boldsymbol{\beta} + W\mathbf{u},$$

where $Z = [Z_{ij}]$, $W = \text{diag}(W_{ij})$, and $\mathbf{u} = [u_i]$. It is assumed as before that $\mathbf{u}_i$ is multivariate normal with mean $\mathbf{0}$ and covariance Σ, $i = 1, \cdots, n$. Also, $\mathbf{e} = [\mathbf{e}_{ij}]$ is a vector of unobserved errors with $E(\mathbf{e}_{ij} \mid \boldsymbol{\pi}_{ij}) = \mathbf{0}$ and $\text{cov}(\mathbf{e}_{ij} \mid \boldsymbol{\pi}_{ij}) = R^{1/2}_{\boldsymbol{\pi}_{ij}} R R^{1/2}_{\boldsymbol{\pi}_{ij}}$. For the multinomial models considered here, the form of $R_{\boldsymbol{\pi}_{ij}}$ is given in Section 2.3. The additional unknown covariance matrix R is included for the modeling of population-average associations. For the complete data model, we define the cov($\mathbf{u}$) $= \mathbf{\Sigma}$ and cov($\mathbf{e} \mid \boldsymbol{\pi}$) $= R^{1/2}_{\boldsymbol{\pi}} \mathbf{R} R^{1/2}_{\boldsymbol{\pi}}$, where $R_{\boldsymbol{\pi}} = \text{diag}(R_{\boldsymbol{\pi}_{ij}})$, and $\mathbf{\Sigma}$ and $\mathbf{R}$ are diagonal matrices with Σ and R on the diagonals, respectively.

The PL procedure is carried out by iteratively fitting a weighted Gaussian linear mixed model to a modified response vector. To achieve this, a number of approximations are required. First, for known estimates of $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$, let the estimated response probabilities be

$$\hat{\boldsymbol{\pi}} = \mathbf{h}(\hat{\boldsymbol{\eta}}) = \mathbf{h}(Z\hat{\boldsymbol{\beta}} + W\hat{\mathbf{u}}),$$

where $\mathbf{h}(\cdot)$ is the response function for the desired model. Then a Taylor series approximation to the residuals $\mathbf{e} = \bar{\mathbf{y}} - \boldsymbol{\pi}$ from (3.74) about the current estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ is given by

$$\begin{aligned} \mathbf{e} \approx \tilde{\mathbf{e}} &= (\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}) - (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}})' \frac{d(\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}})}{d\boldsymbol{\eta}} \\ &= (\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}) - (Z\boldsymbol{\beta} - Z\hat{\boldsymbol{\beta}} + W\mathbf{u} - W\hat{\mathbf{u}})'D, \end{aligned} \tag{3.75}$$

where $D = \text{diag}(D_{ij})$, and D_{ij} is defined as in Chapter 2.

Next, we approximate the conditional distribution of $\tilde{\mathbf{e}} \mid \boldsymbol{\beta}, \mathbf{u}$ in (3.75) with a Gaussian distribution that has the same first two moments as $\mathbf{e} \mid \boldsymbol{\beta}, \mathbf{u}$. Thus we assume that

$$\tilde{\mathbf{e}} \mid \boldsymbol{\beta}, \mathbf{u} \sim MVN(\mathbf{0},\ R_{\boldsymbol{\pi}}^{1/2}\ \mathbf{R}\ R_{\boldsymbol{\pi}}^{1/2}). \tag{3.76}$$

Then, using (3.75), (3.76), and approximating $\boldsymbol{\pi}$ in $R_{\boldsymbol{\pi}}^{1/2}\ \mathbf{R}\ R_{\boldsymbol{\pi}}^{1/2}$ with $\hat{\boldsymbol{\pi}}$, we obtain

$$D^{-1'}(\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}) \sim MVN[Z\boldsymbol{\beta} - Z\hat{\boldsymbol{\beta}} + W\mathbf{u} - W\hat{\mathbf{u}},\ D^{-1}\ R_{\hat{\boldsymbol{\pi}}}^{1/2}\ \mathbf{R}\ R_{\hat{\boldsymbol{\pi}}}^{1/2}\ D^{-1'}]. \tag{3.77}$$

It follows from (3.77) that the approximate “pseudo” observation vector,

$$\check{\mathbf{y}} = \mathbf{g}(\hat{\boldsymbol{\pi}}) + D^{-1'}(\bar{\mathbf{y}} - \hat{\boldsymbol{\pi}}), \tag{3.78}$$

has the approximate conditional distribution

$$\check{\mathbf{y}} \mid \boldsymbol{\beta}, \mathbf{u} \sim MVN[Z\boldsymbol{\beta} + W\mathbf{u},\ D^{-1}\ R_{\hat{\boldsymbol{\pi}}}^{1/2}\ \mathbf{R}\ R_{\hat{\boldsymbol{\pi}}}^{1/2}\ D^{-1'}]. \tag{3.79}$$

Treating $\boldsymbol{\beta}$ as an unknown parameter, (3.79) is of the form of a weighted linear mixed model with response $\check{\mathbf{y}}$ and weight matrix given by $\hat{\mathcal{W}} = D'\ R_{\hat{\boldsymbol{\pi}}}^{-1}\ D$. Note that the modified response $\check{\mathbf{y}}$ in (3.78) is analogous to the pseudo-response defined in Section 2.3 for iteratively re-weighted least squares.

The log-likelihood corresponding to (3.79) is easily obtained. Following Wolfinger and O'Connell (1993), we insert an additional dispersion parameter ϕ into the log-likelihood and re-express the covariance matrices $\boldsymbol{\Sigma}$ and $\mathbf{R}$ as $\boldsymbol{\Sigma}^* = \phi^{-1}\boldsymbol{\Sigma}$ and $\mathbf{R}^* = \phi^{-1}\mathbf{R}$. The dispersion parameter is analogous to that used in quasi-likelihoods and can be forced to be 1.0 if unneeded. The resulting log-likelihood is

$$l(\boldsymbol{\beta}, \phi, \boldsymbol{\Sigma}^*, \mathbf{R}^*) = -\frac{1}{2}\log \mid \phi V \mid -\frac{1}{2}\phi^{-1}(\check{\mathbf{y}} - Z\boldsymbol{\beta})'\ V^{-1}\ (\check{\mathbf{y}} - Z\boldsymbol{\beta}) - \frac{n}{2}\log(2\pi), \tag{3.80}$$

where

$$V = \mathcal{W}^{-1/2}\, \mathbf{R}^*\, \mathcal{W}^{-1/2} + W\, \mathbf{\Sigma}^*\, W'. \tag{3.81}$$

Closed form solutions for $\boldsymbol{\beta}$ and ϕ that maximize (3.80) exist and are given by

$$\hat{\boldsymbol{\beta}} = (Z'\, \hat{V}^{-1}\, Z)^{-1}\, Z\, \hat{V}^{-1} \check{\mathbf{y}}, \tag{3.82}$$

and

$$\hat{\phi} = \frac{1}{n-p}\, \hat{\mathbf{r}}'\, \hat{V}^{-1}\, \hat{\mathbf{r}}, \tag{3.83}$$

where $\hat{\mathbf{r}} = \check{\mathbf{y}} - Z(Z'\, \hat{V}^{-1}\, Z)^{-1}\, Z'\, \hat{V}^{-1}\, \check{\mathbf{y}}$. To obtain the estimates of $\mathbf{\Sigma}^*$ and $\mathbf{R}^*$ in $\hat{V}$, the restricted profile likelihood can be maximized

$$\begin{aligned} l_R(\mathbf{\Sigma}^*, \mathbf{R}^*) = -\frac{1}{2} \log \mid V \mid - \frac{n-p}{2} \log(\mathbf{r}'\, V^{-1}\, \mathbf{r}) - \frac{1}{2} \log \mid Z'\, V^{-1}\, Z \mid \\ - \frac{n-p}{2} \{1 + \log[2\pi/(n-p)]\}. \end{aligned} \tag{3.84}$$

Maximization of (3.84) provides REML estimates for $\mathbf{\Sigma}^*$ and $\mathbf{R}^*$. An estimate of $\mathbf{u}$ can then be found from

$$\hat{\mathbf{u}} = \hat{\mathbf{\Sigma}}^*\, W'\, \hat{V}^{-1}\, \hat{\mathbf{r}}. \tag{3.85}$$

Wolfinger and O'Connell (1993) referred to the PL algorithm that utilized the REML estimation (3.84) as a restricted-PL (REPL) procedure. Using this terminology, the REPL algorithm proceeds as follows:

0. Calculate initial estimates for $\boldsymbol{\beta}^{(0)}$ and $\mathbf{u}^{(0)}$. Set $\mathbf{\Sigma}^{(0)} = \mathbf{R}^{(0)} = I$.

For $s = 0, 1, \cdots$

1. Calculate the modified response $\check{\mathbf{y}}^{(s)}$ using $\boldsymbol{\beta}^{(s)}$ and $\mathbf{u}^{(s)}$.
2. Maximize (3.84) to obtain $\hat{\mathbf{\Sigma}}^*$ and $\hat{\mathbf{R}}^*$. Convert $\hat{\mathbf{\Sigma}}^*$ and $\hat{\mathbf{R}}^*$ to $\mathbf{\Sigma}^{(s)}$ and $\mathbf{R}^{(s)}$ using $\phi^{(s)}$ from (3.83).

3. If change from $(\mathbf{\Sigma}^{(s-1)}, \mathbf{R}^{(s-1)})$ to $(\mathbf{\Sigma}^{(s)}, \mathbf{R}^{(s)})$ is small, then stop. Otherwise, compute $\boldsymbol{\beta}^{(s+1)}$ and $\mathbf{u}^{(s+1)}$ from (3.82) and (3.85) and go to 1.

Upon convergence, an estimate of the approximate covariance matrix for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ can be obtained by inverting

$$\begin{bmatrix} Z' \, \hat{\mathcal{W}}^{1/2} \, \hat{\mathbf{R}}^{-1} \, \hat{\mathcal{W}}^{1/2} \, Z & Z' \, \hat{\mathcal{W}}^{1/2} \, \hat{\mathbf{R}}^{-1} \, \hat{\mathcal{W}}^{1/2} \, W \\ W' \, \hat{\mathcal{W}}^{1/2} \, \hat{\mathbf{R}}^{-1} \, \hat{\mathcal{W}}^{1/2} \, Z & W' \hat{\mathcal{W}}^{1/2} \, \hat{\mathbf{R}}^{-1} \, \hat{\mathcal{W}}^{1/2} \, W + \hat{\mathbf{\Sigma}}^{-1} \end{bmatrix} . \tag{3.86}$$

3.5.3 Discussion

In the motivation of the algorithm in the previous section, we allowed for both a subject-specific covariance matrix and a population-average covariance matrix. For the models discussed in this thesis, only the subject-specific covariance term is considered. Wolfinger and O'Connell (1993) discussed how one might use these matrices individually, but did not elaborate on how one would interpret parameters if both were used in the same model. We also included the overdispersion parameter as proposed by Wolfinger and O'Connell (1993). To allow for comparisons between the exact maximum likelihood methods and the proposed approximate method, we force the overdispersion parameter to be 1.0.

We have noted before that the approximate maximum likelihood methods can perform poorly for binomial responses with small sample sizes. We would suspect that the proposed model will run into similar problems when the multinomial sample size is small. As an example of this, Table 3.1 contains the estimates from both the REPL algorithm and the adaptive Gauss-Hermite algorithm for a dataset taken from Agresti and Lang (1993). The dataset originated from the 1989 General Social Survey in which subjects were asked their opinion on (a) teens having sexual relations before marriage, (b) a man and a women having sexual relations before marriage, and (c) a married person having sexual relations with someone other than their spouse. A total of 475 subjects responded to each of the three questions using the response scale

"Always wrong", "Almost always wrong", "Wrong only sometimes", and "Not wrong at all". The results in Table 3.1 were obtained by fitting a cumulative logit model with a random intercept to account for the correlation between a given subject's responses. For the ith subject and the jth question, the linear predictor has the form

$$\eta_{ijr} = \alpha_r + \beta_1 x_{ij1} + \beta_2 x_{ij2} + u_i, \quad r = 1, \cdots, 4, \quad j = 1, \cdots, 3,$$

where x_{ij1} and x_{ij2} are one if response pertains to teenage or premarital sex, respectively, and zero otherwise. Thus, β_1 and β_2 are the corresponding regression parameters, the $\{\alpha_r\}$ are the threshold parameters, and u_i is assumed to be normally distributed with mean zero and standard deviation σ. We include this example to point out the disparity in the estimates of the standard deviation of the random effect. The estimate from the REPL algorithm is substantially smaller than that of the adaptive quadrature algorithm. The fixed effects parameter estimates and standard errors differ as well, though statistical conclusions from the two models would be same. Note that the log-likelihood values for the two models are not comparable and that, as can be seen from the approximate covariance matrix (3.86), the REPL algorithm does not provide standard error estimates for the variance components.

As has been done for the binary case, further research is needed to investigate the accuracy of the REPL approach for multinomial models. Bias correction terms, such as those proposed by Lin and Breslow (1996), could also be examined. At the very least, the approximate methods considered here can be used to calculate starting values for the exact algorithms considered previously. Their speed and simplicity also makes them ideal for doing exploratory analyses prior to fitting a full "exact" maximum likelihood analysis.

3.6 Applications

We now consider three examples to illustrate the fitting methods discussed in Sections 3.3 and 3.5. Our intent is not to provide a thorough analysis of each dataset,

Table 3.1: Parameter estimates for fitting a cumulative logit model with a random intercept to sexual opinion dataset Agresti and Lang (1993). Results are shown for the 10-point adaptive Gauss-Hermite algorithm (AGH(10)) and the restricted pseudo-likelihood algorithm (REPL)

	AGH(10)	REPL
α_1	2.652	1.922
α_2	3.786	2.890
α_3	5.435	4.267
β_1	-0.571	-0.455
	(.187)	(.172)
β_2	-4.378	-3.340
	(.262)	(.169)
σ	2.267	1.592
	(0.191)	
LL	-1218.429	-7822.265

but to simply apply the methods to some specific models. The data in Table 3.2 are from a wine tasting experiment (Randall 1989) in which wine preferences were measured on an ordinal scale. To account for possible heterogeneity of judges, Tutz and Hennevogl (1996) fit a cumulative logit random intercept model utilizing their EM Gauss-Hermite and EM Monte Carlo algorithms. We fit the same model with our algorithm to allow for comparisons between the approaches. We also illustrate our methods using an adjacent-category logit random intercept model. The second dataset in Table 3.3 arose from a developmental toxicity study using litters of mice conducted under the U.S. National Toxicology Program (Price et al. 1985). The three possible outcomes in the study (Dead/Resorption, Malformation, Normal) have a natural sequential ordering which lends itself to a continuation-ratio logit model. We incorporate random effects to account for the correlations between fetuses in the same litter. The final dataset, Table 3.4, is from the 1975 U.S. General Household Survey in which subjects indicated their degree of satisfaction with family (F), hobbies (H),

Table 3.2: Bitterness of wine data (Randall 1989) classified by temperature, presence or absence of skin contact, and bottle number.

	Low Temperature				High Temperature			
	No Contact		Contact		No Contact		Contact	
Judge	Bottle 1	Bottle 2	Bottle 1	Bottle 2	Bottle 1	Bottle 2	Bottle 1	Bottle 2
1	2	3	3	4	4	4	5	5
2	1	2	1	3	2	3	5	4
3	2	3	3	2	5	5	4	4
4	3	2	3	2	3	2	5	3
5	2	3	4	3	3	3	3	3
6	3	2	3	2	2	4	5	4
7	1	1	2	2	2	3	2	3
8	2	2	2	3	3	3	3	4
9	1	2	3	2	3	2	4	4

and residence (R) (Clogg 1979). Such item response data is common in psychometric literature. Hedeker (2000) utilized a baseline-category logit model with random effects to analyze Table 3.4. To allow for comparisons, we fit similar random effects models. We note that all results and computing times reported were obtained on a Sun Enterprise 450 computer which had four 400 MHz. processors and one gigabyte of RAM.

3.6.1 Wine Tasting Experiment

Table 3.2 contains the data from a study on the bitterness of white wine (Randall 1989). Of interest in the study was whether certain factors that can be controlled during the pressing of the grapes influenced the bitterness of the wine. The factors considered were the temperature during pressing and whether there was contact of the juice with the skin when the grapes were crushed. Temperature was considered as either high or low and contact was measured by presence or absence. At each of the temperature/contact combinations, two bottles of white wine were randomly chosen and the bitterness of each was classified on a five-point ordinal scale from least to most bitter. For this factorial experiment, nine professional judges were chosen to

Table 3.3: Developmental toxicity data (Price et al. 1985) classified by ethylene glycol dosage and frequency of fetus outcome (D/R = Dead/Resorption, M = Malformation, N = Normal).

	Ethylene Glycol Dosage (g/kg)											
	0.00			0.75			1.50			3.00		
Litter	D/R	M	N	D/R	M	N	D/R	M	N	D/R	M	N
1	1	0	7	0	3	7	0	8	2	0	4	3
2	0	0	14	1	3	11	0	6	5	1	9	1
3	0	0	13	0	2	9	0	5	7	0	4	8
4	0	0	10	0	0	12	0	11	2	1	11	0
5	0	1	15	0	1	11	1	6	3	0	7	3
6	1	0	14	0	3	10	0	7	6	0	9	1
7	1	0	10	0	0	15	0	0	1	0	3	1
8	0	0	12	0	0	11	0	3	8	0	7	0
9	0	0	11	2	0	8	0	8	3	0	1	3
10	0	0	8	0	1	10	0	2	12	0	12	0
11	1	0	6	0	0	10	0	1	12	2	12	0
12	0	0	15	0	1	13	0	10	5	0	11	3
13	0	0	12	0	1	9	0	5	6	0	5	6
14	0	0	12	0	0	14	0	1	11	0	4	8
15	0	0	13	1	1	11	0	3	10	0	5	7
16	0	0	10	0	1	9	0	0	13	2	3	9
17	0	0	10	0	1	10	0	6	1	0	9	1
18	1	0	11	0	3	10	0	2	6	0	0	9
19	0	0	12	0	0	15	0	1	2	0	5	4
20	0	0	13	0	0	15	0	0	7	0	2	5
21	1	0	14	0	2	5	0	4	6	1	3	9
22	0	0	13	0	1	11	0	0	12	0	2	5
23	0	0	13	0	1	6				0	1	11
24	1	0	14	1	1	8						
25	0	0	14									

Table 3.4: 1975 U.S. General Household Survey data (Clogg 1979) concerning degree of satisfaction with family (F), hobbies (H), and residence (R) on a three-point scale (1=Low, 2=Medium, 3=High).

Response Profile				Response Profile				Response Profile			
F	H	R	Frequency	F	H	R	Frequency	F	H	R	Frequency
1	1	1	15	1	2	1	3	1	3	1	5
1	1	2	11	1	2	2	12	1	3	2	14
1	1	3	7	1	2	3	5	1	3	3	16
2	1	1	16	2	2	1	23	2	3	1	18
2	1	2	26	2	2	2	58	2	3	2	38
2	1	3	12	2	2	3	31	2	3	3	27
3	1	1	23	3	2	1	45	3	3	1	64
3	1	2	49	3	2	2	117	3	3	2	191
3	1	3	54	3	2	3	126	3	3	3	466

rate each of the eight bottles of wine. Since the judges cannot be expected to have the same sensitivity to the bitterness of wine, one would expect their individual ratings to be correlated. An analysis of this data should account for the heterogeneity of judges.

To account for the heterogeneity, Tutz and Hennevogl (1996) fit a cumulative logit random effects model that allowed the judges to have shifted thresholds. They fit a random intercept model that included factors for temperature, contact, and bottle. That is for the jth evaluation by the ith judge,

$$\eta_{ijr} = \alpha_r + \beta_{TE}\, x_{ij1} + \beta_{CO}\, x_{ij2} + \beta_{BO}\, x_{ij3} + u_i, \qquad (3.87)$$
$$r = 1, \cdots, R-1,\ j = 1, \cdots, T,\ i = 1, \cdots, n,$$

where $R = 5$, $T = 8$, and $n = 9$. In (3.87) β_{TE}, β_{CO}, and β_{BO} are the parameter coefficients for the temperature, contact, and bottle factors, respectively. Tutz and Hennevogl (1996) utilized effect coding (effects sum to zero) for the covariates. Following their coding scheme, x_{ij1} = temperature was coded as (Low=1, High=-1),

Table 3.5: Parameter estimates and log-likelihood values (LL) for fitting model (3.87) to the wine tasting dataset using the cumulative logit link. Numbers in column labels denote the number of quadrature points or Monte Carlo samples used in the adaptive Gauss-Hermite (AGH), Gauss-Hermite EM (GHEM) (Tutz and Hennevogl 1996), and Monte Carlo EM (MCEM) (Tutz and Hennevogl 1996) algorithms. MCEM_A refers to the automated Monte Carlo EM algorithm, REPL refers to the restricted pseudo-likelihood algorithm, and FIXED refers to the fixed effects model obtained by omitting the random effect.

	FIXED	AGH(5)	MCEM_A	REPL	GHEM(10)	MCEM(20)
α_1	-3.359	-4.082	-4.057	-3.993	-4.139	-4.439
α_2	-0.762	-0.930	-0.923	-0.911	-0.969	-1.234
α_3	1.456	1.797	1.787	1.755	1.777	1.519
α_4	2.994	3.657	3.638	3.585	3.649	3.368
β_{TE}	1.251	1.536	1.527	1.501	1.546	1.549
	(.264)	(.298)	(.295)	(.287)	(.437)	(1.092)
β_{CO}	0.763	0.916	0.911	0.894	0.925	0.925
	(.238)	(.256)	(.257)	(.248)	(.347)	(.825)
β_{BO}	0.048	0.122	0.120	0.120	0.123	0.126
	(.223)	(.232)	(.236)	(.228)	(.320)	(.753)
σ	–	1.145	1.105	1.213	1.243	1.261
		(.401)	(.397)	–	(.479)	(.954)
LL	-86.469	-81.394	–	–	-81.365	-80.437

x_{ij2} = contact was coded as (No Contact = 1, Contact = -1), and x_{ij3} = bottle was coded as (Bottle 1 = 1, Bottle 2 = -1).

In Table 3.5 are the results of fitting model (3.87) with the adaptive Gauss-Hermite quadrature, Monte Carlo, and REPL algorithms given in Sections 3.3 and 3.5. Included in Table 3.5 are the results reported by Tutz and Hennevogl (1996) as well as the estimates obtained from fitting the fixed effects model. To determine the number of quadrature points for the adaptive Gauss-Hermite algorithm, the number of nodes was successively increased until the difference in parameter and standard error estimates between successive fits was less than 0.0001. Five quadrature nodes were found to be sufficient to obtain the desired accuracy. The algorithm required less than 30 seconds to obtain the final parameter estimates using five quadrature nodes. The automated MCEM algorithm, starting from the REPL final estimates, took approximately 23 hours to obtain convergence starting from a simulation size of 100 and ending at 99,365. As noted before, the MCEM algorithm is useful for models with high numbers of random effects but is generally inefficient for low to moderate sized models when compared with adaptive Gauss-Hermite quadrature. The REPL algorithm required less than 30 seconds to obtain convergence. Due to the large matrices (the design matrix Z for the entire data has dimensions 288 by 7), it was necessary to fit the model by looping over clusters instead of using the entire data directly as in the definition of the algorithm in Section 3.5. The final two columns in Table 3.5 contain the estimates reported by Tutz and Hennevogl (1996) using their EM algorithms.

Examining the AGH(5) column in Table 3.5, it is clear that both temperature and contact impact the perceived bitterness of the wine. The Wald statistics for these two factors are 26.5 and 12.8, respectively, which are highly significant ($P < 0.001$). The positive sign of both the temperature and contact coefficients indicates that the lower temperature and no contact levels are associated with lower perceived bitterness. For example, holding bottle and contact fixed the odds of bitterness being below any fixed level is $\exp(2 * 1.536) = 21.6$ times greater for low temperature than for high temperature for a given subject. The estimated standard deviation of the random effect is 1.15, indicating that the judges did indeed vary with respect to their perceived bitterness. Though similar inferential results are obtained from both the fixed and random effects models, the parameter estimates and standard errors in the random effects model are correctly adjusted for the unobserved heterogeneity of the judges.

The results of the five different algorithms for fitting the random effects model are generally in agreement. The adaptive MCEM and REPL results are very similar to the adaptive quadrature results, with the latter approach providing a slightly larger estimated standard deviation for the random effect. The parameter estimates from the EM algorithms of Tutz and Hennevogl (1996) are generally larger than those obtained using adaptive quadrature. In addition, the standard errors are much larger than those estimated by the AGH, automated MCEM, and REPL methods. Recall that the approach for finding standard errors used by Tutz and Hennevogl (1996) has a tendency to overestimate the true standard errors. However, the extremely large standard errors for their MCEM algorithm (nearly three times those of the adaptive quadrature method) are due in part to an error in their programming (personal communication, Tutz). Though not reported in Table 3.5, we also fit model (3.87) using the direct, Gauss-Hermite maximization approach of Hedeker and Gibbons (1994). Using their approach, we needed 25 quadrature points to obtain the same results that were obtained using five point adaptive quadrature.

For this particular dataset, one might also be interested in predicting the individual judge effects. Such predictions can be obtained using the methods described in Section 3.4.3 or through the approximate REPL procedure. Table 3.6 contains the predicted judge effects using adaptive quadrature, the REPL procedure, and the estimates reported by Tutz and Hennevogl (1996) using Gauss-Hermite quadrature. We also include the estimates obtained by treating judge as a fixed effect. It is clear from the fixed effect estimates that the judges have varying scales of bitterness with judges 1 (-2.42) and 7 (2.83) begin quite different from each other. The predicted estimates obtained from adaptive quadrature maintain the same ordering as in the fixed effects case, but are smaller in magnitude. For example the thresholds for judge 7 are predicted to be shifted by 1.93 from the overall thresholds reported in Table 3.5. Such smoothing of the estimates is common as the random effect approach borrows information across all judges to obtain the predictions. The approximate REPL approach provides very similar predictions for the random effects. We again see that the

Table 3.6: Predicted judge effects from the cumulative logit model (3.87) with standard errors. Numbers in column labels denote the number of quadrature points used in the adaptive Gauss-Hermite (AGH) or Gauss-Hermite EM (GHEM) (Tutz and Hennevogl 1996) algorithms. REPL refers to the restricted pseudo-likelihood algorithm. FIXED refers to the fixed effects model obtained by omitting the random effect.

Judge	FIXED	AGH(5)	REPL	GHEM(10)
1	-2.423 (.716)	-1.717 (.759)	-1.731 (.704)	-1.853 (.505)
2	0.905 (.666)	0.598 (.710)	0.608 (.694)	0.650 (.563)
3	-1.480 (.670)	-0.992 (.725)	-1.007 (.694)	-0.952 (.622)
4	0.078 (.659)	0.052 (.693)	0.055 (.692)	0.101 (.686)
5	-0.357 (.659)	-0.234 (.687)	-0.236 (.691)	-0.261 (.648)
6	-0.713 (.660)	-0.473 (.689)	-0.478 (.692)	-0.498 (.560)
7	2.830 (.769)	1.929 (.816)	1.954 (.719)	2.021 (.494)
8	0.380 (.660)	0.273 (.653)	0.276 (.692)	0.395 (.567)
9	0.780 (.673)	0.552 (.672)	0.559 (.693)	0.621 (.484)

GHEM algorithm of Tutz and Hennevogl (1996) provides generally larger estimates than those obtained using adaptive quadrature. Standard errors for the adaptive quadrature procedure are based on an approximation to conditional mean square error of prediction described by Booth and Hobert (1998). Though they did not report how they calculated the standard errors, it appears that Tutz and Hennevogl (1996) did not adjust for the variability associated with using estimates for the model parameters, which resulted in smaller standard error estimates.

Instead of using a cumulative logit link for modeling Table 3.2, one could also utilize the adjacent-category link. The effects in the cumulative logit model refer to the entire response scale. In contrast, the effects in the adjacent-category logit model refer to the multiplicative effect of a one-unit increase of a predictor on the odds of response in the higher instead of the lower of any two adjacent categories. Generally both models will fit well in similar situations as they both imply stochastic orderings of the response distributions for different predictor values. Thus the choice of one link over another will depend on the desired interpretation.

Using the appropriate design matrix and link function from Section 2.4.2, we fit model (3.87) using the adjacent-category logit link. Results are shown in Table 3.7 for the adaptive quadrature and REPL algorithms. For the adaptive quadrature algorithm, eight quadrature points were required to obtain the desired accuracy. The inferential results are substantively the same for the adjacent-category logit model as for the cumulative logit model. We again see that both temperature and contact are associated with the perceived bitterness. Holding all other factors constant and for a given subject, the odds for the low temperature that bitterness value is 2 instead of 1 (or 3 instead of 2, or 4 instead of 3, or 5 instead of 4) is estimated to be $\exp(2 * (-1.149)) = 0.10$ times the odds for the high temperature. To parallel the cumulative logit results, the odds of 1 instead of 2 is $\exp(2 * 1.149) = 9.95$ times greater for the low than the high temperature. The estimated odds are greater for the cumulative logit model (e.g. 21.6 for temperature) as that link refers to the entire response scale. The estimated standard deviation for the random effect is 0.84 for the adaptive quadrature procedure with similar results for the REPL procedure. Again we see that effects and standard errors are larger than the corresponding fixed effects estimates when the heterogeneity of the judges is taken into account.

3.6.2 Developmental Toxicity Data

Table 3.3 displays the results of a toxicity study involving ethylene glycol. Such experiments are used to test and regulate substances which may pose potential danger to developing fetuses. In this experiment pregnant mice were randomly assigned to one of four dosages groups (0, 0.75, 1.50, 3.00 g/kg) of ethylene glycol. Each group of mice was exposed to the ethylene glycol concentration and then their fetuses were examined for defects. Each fetus was classified as either Dead/Resorption, Malformed, or Normal. The continuation-ratio logit is a natural link function for this type of response as obtaining a given classification is dependent on passing through the prior classifications. A random effects approach is also warranted for this type of

Table 3.7: Parameter estimates for fitting model (3.87) using the adjacent-category logit model. Numbers in column labels denote the number of quadrature points used in the adaptive Gauss-Hermite (AGH) algorithm. REPL refers to the restricted pseudo-likelihood algorithm and FIXED refers to the fixed effects model obtained by omitting the random effect.

	FIXED	AGH(8)	REPL
α_1	0.026	-0.009	-0.001
α_2	-0.738	-1.043	-0.953
α_3	-1.326	-1.880	-1.716
α_4	-1.461	-2.250	-2.017
β_{TE}	-0.845	-1.149	-1.058
	(.205)	(.268)	(.234)
β_{CO}	-0.483	-0.659	-0.607
	(.162)	(.203)	(.184)
β_{BO}	-0.041	-0.056	-0.051
	(.143)	(.167)	(.160)
σ	–	0.839	0.832
		(0.193)	
LL	-85.603	-80.853	–

data as fetuses within the same litter are likely to be correlated due to common genetic factors passed on by the dam. For instance, some dams may be more susceptible to the ethylene glycol which would then be inherited by the fetuses.

Coull and Agresti (2000) analyzed this dataset using a multivariate binomial logit-normal (BLN) model. The multivariate BLN model is a generalization of the logit-normal model that allows one to model the correlation structure among a set of binomial response variables. In contrast to the logit-normal model which allows only positive correlations among observations from the same cluster, the multivariate BLN model allows for a wide variety of covariance structures for modeling the observations within a cluster. Coull and Agresti (2000) exploited the fact that the multinomial mass function can be factored into a product of binomial mass functions. Thus the multinomial response for a model using the continuation-ratio logit can be treated a set of independent binomial counts. This allowed them to fit a continuation-ratio logit random effects model using their multivariate BLN model. To evaluate the intractable integrals, Coull and Agresti (2000) employed Gauss-Hermite quadrature with 20 quadrature points. We fit similar models from the framework of a multinomial random effects model using adaptive Gauss-Hermite quadrature.

We begin by fitting a simple model that assumes a common dosage effect for the two logits. The two logits model the probability of a dead/resorbed fetus and the conditional probability of a malformed fetus given the fetus was alive. To account for the possibility of litter effects, we include a litter-specific random effect which allows the logits for each litter to be shifted. The linear predictor for the ith litter has the form

$$\eta_{ir} = \alpha_r + \beta_{DO}\, x_i + u_i, \quad r = 1, \cdots, R-1, \quad i = 1, \cdots, n, \tag{3.88}$$

where $R = 3$, $n = 94$, and β_{DO} denotes the parameter for the dosage predictor. Table 3.8 displays the results for this model using the adaptive quadrature, automated MCEM, and REPL algorithms. The adaptive quadrature algorithm required 12 quadrature points and converged in less than 30 seconds. Starting from the REPL estimates, the automated Monte Carlo EM algorithm required 43 minutes to converge with a final simulation sample size of 2,361. The REPL algorithm required less than 30 seconds to converge. Interpreting the adaptive quadrature results, the estimated odds of death and the estimated odds of malformation given survival are multiplied by $\exp(1.303) = 3.68$ for every additional g/kg of ethylene glycol. Accounting for the heterogeneity of the litters increased both the dosage effect size and its standard error when compared with the fixed effects analysis. The variation among the litters was fairly large with an estimated standard deviation of 1.18. The Monte Carlo EM algorithm provided similar estimates to the adaptive quadrature approach. The approximate REPL algorithm tended to underestimate the parameters. This may be due to the large number of clusters (94). The estimating equations in the REPL procedure simultaneously estimate the fixed effect parameters and the random effect parameters. Such a large number of random effect parameters may adversely effect the fixed effect parameter estimates.

Model (3.88) makes the strong assumption that the ethylene glycol dosage influences the probability of death and the conditional probability of malformation given survival in the same manner. This assumption can be relaxed by allowing a separate dosage effect parameter for each logit. The shifting of thresholds assumption in model (3.88) can also be relaxed by allowing the probability of death and the conditional probability of malformation given survival to vary individually. The variation in probabilities may be, for example, due to differing susceptibilities of the mice to the dosages. This is accomplished by replacing u_i in (3.88) with a litter-logit-specific

Table 3.8: Parameter estimates for fitting model (3.88) using the continuation-ratio logit link. Numbers in column labels denote the number of quadrature points used in the adaptive Gauss-Hermite (AGH) algorithm. MCEM$_A$ denotes the automated Monte Carlo EM algorithm, while REPL refers to the restricted pseudo-likelihood algorithm. The fixed effects results (FIXED) were obtained by fitting model (3.88) without a random effect.

	FIXED	AGH(12)	MCEM$_A$	REPL
α_1	-5.852	-7.020	-7.008	-6.450
α_2	-2.707	-3.398	-3.392	-3.052
β_{DO}	1.041	1.303	1.301	1.176
	(.073)	(.135)	(.135)	(.134)
σ	–	1.175	1.157	1.058
		(0.153)		
LL	-539.765	-494.528	–	–

random effect u_{ir} yielding the model

$$\eta_{ir} = \alpha_r + \beta_{DO_r}\, x_i + u_{ir}. \tag{3.89}$$

This new model has a multivariate random effect $\mathbf{u}_i = (u_{i1}, u_{i2})'$ which we assume to be multivariate normal with zero mean and covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$. In addition to the fixed effects version of model (3.89), Table 3.9 contains the results for three variations of model (3.89) using the adaptive quadrature algorithm. In the second column of results we fit model (3.89) allowing only shifted thresholds, as in model (3.88), with $\sigma_1 = \sigma_2$ and $\rho = 1$. We then allow for separate random effects for each logit, but assume that the random effects are independent. That is $\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$. The estimates from this model are found in the third column of results. For the final column of results we allow the random logit effects to be correlated.

From all four models it is obvious that there is a dosage effect on malformation given survival but not on death. The models allowing varying logits provide very similar results. One could perform a likelihood-ratio test for comparing these two models with a null hypothesis of $H_o : \rho = 0$. One would conclude that the model without the correlation is sufficient as the likelihood-ratio test statistic is only .022. Comparisons of the shifted threshold model with the varying threshold models would require a score test as the null hypothesis model would contain parameters on the boundary of the parameter space ($\sigma_2^2 = 0$). We consider such a test in Chapter 5. From the log-likelihood values alone, one would conclude that the varying threshold model without correlation is the most appropriate model. It is evident from the standard deviation estimates for this model that the litter effect is much stronger for the malformation given survival outcome than the death outcome. Coull and Agresti (2000) considered a number of other models that allowed different covariance structures for each dosage group. They also concluded that the varying threshold model without correlation described the data well.

3.6.3 1975 General Household Survey Life Satisfaction Data

The final dataset, Table 3.4, comes from the 1975 U.S. General Household Survey. A total of 1,472 subjects were asked to rate their degree of life satisfaction with their family, hobbies, and residence using a three-point scale (1 = Low, 2 = Medium, 3 = High). A number of different approaches have been used to analyzed this particular dataset, such as a 3-class latent variable model Clogg (1979) and a latent trait model Masters (1985), which are summarized in Bartholomew (1987). Recently, Hedeker (2000) re-analyzed the survey data using a baseline-category logit random effects model. A nominal logistic regression model expresses the responses in terms of two logits, namely, $\log \frac{P\{High\}}{P\{Low\}}$ and $\log \frac{P\{Med\}}{P\{Low\}}$. Hedeker (2000) demonstrated his modeling approach by fitting a number of models similar to those found in Bartholomew (1987). As noted before, however, his model does not allow estimation

Table 3.9: Parameter estimates for various fits of the continuation-ratio logit model (3.89) to the toxicity dataset using the adaptive Gauss-Hermite algorithm. Q. PTS. denotes the number of quadrature points used.

		SHIFTED	VARYING LOGITS	
	FIXED	LOGITS	$\rho = 0$	$\rho \neq 0$
Q. PTS.	–	12	18	18
α_1	-4.058	-4.525	-4.196	-4.198
α_2	-2.949	-3.911	-4.360	-4.356
β_{DO_1}	0.094	-0.131	0.083	0.083
	(.200)	(.205)	(.217)	(.217)
β_{DO_2}	1.179	1.588	1.781	1.780
	(.080)	(.160)	(.220)	(.219)
σ_1	–	1.340	0.559	0.559
σ_2	–	–	1.586	1.587
ρ	–	–	0.000	0.080
LL	-526.908	-473.977	-464.744	-464.733

of correlations between random effects in different thresholds. We analyze Table 3.4 using similar models to that of Hedeker (2000) but allowing for correlated random effects between thresholds.

Table 3.4 contains 1,472 subjects or clusters. The exact maximum likelihood algorithms in Section 3.3 are carried out by numerically (by quadrature or sampling methods) evaluating the integrals for each cluster. To directly fit Table 3.4 using this approach would take an exorbitant amount of time, especially when including multiple random effects. One can cleverly "reduce" the number of clusters by noting that there are only $3^3 = 27$ distinct response patterns possible from the survey. All subjects with the same response pattern, e.g. the 15 subjects with response 111, will contribute the same amount to the calculation of the marginal log-likelihood. Thus, within a given iteration, the numerical approximation of the integrals only needs to be carried out for the 27 distinct response profiles. The 27 contributions to the marginal log-likelihood are then multiplied by the appropriate numbers of subjects who responded in that manner. Such a modification is straightforward in the adaptive Gauss-Hermite quadrature algorithm as the marginal log-likelihood is directly maximized. Thus, the results given below were obtain using adaptive quadrature. This also allows us to compare the direct Gauss-Hermite maximization algorithm of Hedeker (2000) to our direct adaptive Gauss-Hermite algorithm.

To analyze Table 3.4 we fit a series of models that are similar to those used in the continuation-ratio logit example. For the jth question of the ith subject, let x_{ij1}, x_{ij2}, and x_{ij3} be indicator variables for the family, hobbies, and residence items, respectively, such that $x_{ijl} = 1$ if $j = l$, $j, l = 1, \cdots, 3$, and zero otherwise. As in Hedeker (2000), we consider models that allow separate item parameters for each logit, which we denote, for the rth logit, by β_{F_r}, β_{H_r}, and β_{R_r} for the family, hobbies, and residence items, respectively. We first fit a simple fixed effects model

$$\eta_{ijr} = \beta_{F_r}\, x_{ij1} + \beta_{H_r}\, x_{ij2} + \beta_{R_r}\, x_{ij3}, \quad r = 1, 2, \quad n = 1, \cdots, 1472. \tag{3.90}$$

Note that in order to directly estimate the three item parameters for each logit, we removed the threshold parameters $\{\alpha_r\}$ found in the original definition of the baseline-category logit model (see Section 2.4.1). Model (3.90) is unrealistic as responses from the same subject are certain to be correlated. We then allowed for a shift in thresholds by including a subject-specific intercept term. This model assumes that the random subject-effect influences the logits High versus Low and Medium versus low in the same manner. Such an assumption is inappropriate for the baseline-category logit model since the nominal responses need not have any relation to each other. Thus one should not expect the subject-specific effect to remain the same across all logits. More realistically, one could allow for varying subject effects for the two logits, as was done in the continuation-ratio logit model. This model has the form

$$\eta_{ijr} = \beta_{F_r}\, x_{ij1} + \beta_{H_r}\, x_{ij2} + \beta_{R_r}\, x_{ij3} + u_{ir}, \tag{3.91}$$

where u_{ir} is a threshold-specific random effect. Recall that the first logit compares Medium versus Low while the second compares High versus Low.

The specification of model (3.91) is not complete until we specify the structure of the covariance matrix $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ for the random effects vector $\mathbf{u}_i = (u_{i1}, u_{i2})'$. In the baseline-category logit random effects model proposed by Hedeker (2000), the covariance term is forced to be $\sqrt{\sigma_1^2 \sigma_2^2}$ so that the correlation ρ is equal to one. This assumption is a consequence of the estimation algorithm he used. His general model has the form

$$\eta_{ijr} = \mathbf{z}_{ij}^{'}\boldsymbol{\beta}_r + \mathbf{w}_{ij}^{'}\mathbf{u}_{ir},$$

where $\mathbf{z}_{ij}$ and $\mathbf{w}_{ij}$ are the fixed and random effects design vectors, respectively, $\boldsymbol{\beta}_r$ are the q fixed effects parameter vectors, and $\mathbf{u}_{ir}$ is the cluster-logit-specific random effects vector for the rth logit. Additionally he assumed that the random effects vector $\mathbf{u}_{ir}$ followed a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ_r. Note that the covariance matrix is allowed to vary at the logit level r. To simplify the use of multivariate Gauss-Hermite quadrature, Hedeker (2000) standardized the random effects by letting $\mathbf{u}_{ir} = \Sigma_r^{1/2}\boldsymbol{\theta}_i$, where $\Sigma_r^{1/2}$ is a matrix containing the elements of the lower Cholesky square root of Σ and $\boldsymbol{\theta}_i$ is a multivariate standard normal random variable. The resulting model has the form

$$\eta_{ijr} = \mathbf{z}_{ij}^{'}\boldsymbol{\beta}_r + \mathbf{w}_{ij}^{'}\Sigma_r^{1/2}\boldsymbol{\theta}_i.$$

To estimate the regression parameter vectors $\boldsymbol{\beta}_r$ and the Cholesky elements of $\Sigma_r^{1/2}$, Hedeker (2000) used a Fisher scoring algorithm on the marginal likelihood obtained through Gauss-Hermite quadrature.

Since the covariance matrix Σ_r is defined at the logit level, the distribution of the random effects for cluster i has a block diagonal covariance matrix $\Sigma = [\Sigma_r]$ with zeros above and below the block diagonal. Thus the random effects terms between logits are assumed to be independent. However, in the Fortran program MIXNO (Mixed-effects Nominal Logistic Regression) that Hedeker (2000) made available to implement his methods, this is not how the model is fit. In MIXNO the random effects vector and, in turn, the covariance matrix are defined at the cluster level as in our model. Thus, for this example, the MIXNO model allowing a shifting in thresholds would be

$$\eta_{ijr} = \beta_{F_r}\ x_{ij1} + \beta_{H_r}\ x_{ij2} + \beta_{R_r}\ x_{ij3} + \sigma\theta_i,$$

where θ_i is a standard normal random variable. To allow for varying thresholds, however, MIXNO actually fits, in contrast to (3.91), the model

$$\eta_{ijr} = \beta_{F_r}\ x_{ij1} + \beta_{H_r}\ x_{ij2} + \beta_{R_r}\ x_{ij3} + \sigma_1\theta_i I[r=1] + \sigma_2\theta_i I[r=2],$$

where $I[r = l]$, $l = 1, 2$, is an indicator function which is one if $r = l$ and zero otherwise. Note that the two random effects, u_{i1} and u_{i2}, in (3.91), have been replaced by a single random effect θ_i. Thus, for our varying threshold model (3.91) to be equivalent with that fit by MIXNO, we would have to assume that u_{i1} and u_{i2} are perfectly (positively) correlated. For the ith subject, let (ξ_{i1}, ξ_{i2}) be the two realizations of the random threshold effects for model (3.91). The realization ξ_{ir} is the amount that the threshold for the ith subject is perturbed from the overall mean threshold α_r, i=1,2. The assumption of perfect correlation between the threshold random effects implies that for all subjects the realizations (ξ_{i1}, ξ_{i2}), $i = 1, \cdots, n$, lie on a line in $\mathbb{R}^2$, with positive slope. Thus, the thresholds for a given subject can vary from the overall mean of the thresholds, but the amounts that they are perturbed will be linearly related to the perturbations for all other subjects. As the amounts that the thresholds are perturbed are related to a subject's perception of utility for the choices, it seems inappropriate to assume that the perception of utility across the choices for all subjects will follow such a perfect relationship. Indeed, if the nominal responses referred to choice of political party (i.e., democrat, republican, independent), one would expect quite different utility perceptions for subjects with differing political ideologies.

The same approach is used by MIXNO to allow random covariate effects to vary across logits. The reason that this approach is used is that it greatly reduces the number of integrals that need to be approximated. Instead of, in general, having a set of q integrals for each cluster, there is only one set of integrals to approximate. The consequence of this approach, however, is that random effects between logits are always perfectly correlated. In Table 3.10 are the results of fitting the fixed effects model (3.90), and various version of the random effects model (3.91) with our method and that proposed by Hedeker (2000).

For all models in Table 3.10, including the fixed effects model, the item fixed effects parameters are highly significant indicating the increased probability of medium or high satisfaction responses, relative to low satisfaction, for all items. Allowing for a shift in thresholds increased both the standard errors and effects sizes, with the estimated standard deviation of the random effect being 1.14. This trend held for the other models as well, except for the varying threshold model with the correlation fixed at zero. We included this model as it coincides with the original model definition given by Hedeker (2000). Though the log-likelihood value for this model is smaller than the shifted threshold model, the parameter estimates are quite different from the remaining varying threshold models. In fact, they are very similar to those obtained for the fixed effects model, especially for the first logit comparing medium versus low. The second to last column contains the results for the varying threshold model with the correlation fixed at 1.0. These results were obtained using MIXNO and are reported in Hedeker (2000). Comparing these estimates to the final column where the correlation was estimated, we see that they are in better agreement than

Table 3.10: Parameter estimates for various fits of the baseline-category logit model (3.91) to the life satisfaction dataset using the adaptive Gauss-Hermite algorithm (AGH) and the Gauss-Hermite algorithm (GH) of Hedeker (2000). The numbers in the column labels denote the number of quadrature points used.

		SHIFTED	VARYING LOGITS		
	FIXED	LOGITS	$\rho = 0$	$\rho = 1$	$\rho = \hat{\rho}$
		AGH(13)	AGH(15)	GH(20)	AGH(15)
β_{F_1}	1.040	1.572	1.004	1.327	1.384
	(.124)	(.161)	(.128)	(.161)	(.169)
β_{H_1}	0.679	1.083	0.658	0.882	0.933
	(.084)	(.116)	(.074)	(.110)	(.119)
β_{R_1}	0.890	1.295	0.880	1.074	1.144
	(.082)	(.114)	(.084)	(.104)	(.115)
β_{F_2}	2.557	3.089	2.949	3.166	3.264
	(.111)	(.151)	(.127)	(.150)	(.161)
β_{H_2}	1.371	1.775	1.477	1.615	1.709
	(.077)	(.111)	(.091)	(.110)	(.118)
β_{R_2}	1.256	1.661	1.276	1.403	1.509
	(.078)	(.112)	(.091)	(.109)	(.118)
σ_1	–	1.142	0.442	0.333	0.832
σ_2	–	–	1.320	1.582	1.626
ρ	–	–	0.000	1.000	0.617
LL	-3854.96	-3828.93	-3744.66	-3742.23	-3736.93

the model with $\rho = 0$. Thus allowing for some correlation, albeit a correlation of 1.0, does provide estimates that are closer to the most general model. However, we do see some changes in the parameter estimates when the correlation is allowed to be estimated. For one, we see that the estimated standard deviation for the first logit (.832) is more than doubled than when $\rho = 1$ (.333). Also, for all of the item parameters, the effect sizes and standard errors are larger.

In practice, on would be interested in comparing the estimated item parameters between the two logits. For example, for the varying threshold model allowing correlation between thresholds, the odds of a subject having medium versus low satisfaction in their family is $\exp(1.384 - 0.933) = 1.6$ times greater than having medium versus low satisfaction in their hobbies. In contrast, the odds for a given subject of choosing high versus low satisfaction for family is $\exp(3.264 - 1.709) = 4.7$ times greater than choosing high versus low satisfaction for hobbies. A similar pattern occurs when comparing family satisfaction versus residence satisfaction (odds of $\exp(1.384 - 1.144) = 1.3$ and $\exp(3.264 - 1.509) = 5.8$ for medium versus low and high versus low, respectively). The comparison of satisfaction with hobbies and satisfaction with residence yields similar odds for medium versus low ($\exp(0.933 - 1.144) = 0.8$) and high versus low ($\exp(1.709 - 1.509) = 1.2$). Thus one might consider fitting a simpler model which constrains the hobbies and residence parameters to be the same within each logit (i.e., $\beta_{H_r} = \beta_{R_r}$, $r = 1, 2$). In summary, the main effect observed is a greater tendency for high satisfaction with families than with hobbies or residence.

3.7 Cumulative Logit Models with Random Thresholds

Recall the wine tasting dataset of Table 3.2. To account for possible differences in sensitivity to wine bitterness among the judges, we analyzed the data in Section 3.6.1 with a cumulative logit model that allowed the thresholds to be randomly shifted for each judge. An underlying assumption in the shifted threshold model is that all thresholds in a given cluster are shifted by the same amount. This assumption also implies that the distance between thresholds remains the same across all clusters. As an example, the estimated thresholds for the wine dataset were $\alpha_1 = -4.082$, $\alpha_2 = -.930$, $\alpha_3 = 1.797$, and $\alpha_4 = 3.657$ (Table 3.5, AGH(5) results). These estimates represent the average threshold estimates across all judges. The judge-specific thresholds are shifted from the average thresholds by the predicted judge effects given in Table 3.6. For each judge, however, the distance between thresholds will remain $\alpha_4 - \alpha_3 = 1.860$, $\alpha_3 - \alpha_2 = 2.727$, and $\alpha_2 - \alpha_1 = 3.152$. Tutz and Hennevogl (1996) proposed a cumulative logit random effects model which relaxed this condition. For a given cluster, they assumed that each threshold was allowed to vary according to its own distribution. Thus the distance between thresholds could vary across judges.

We have already considered models in the previous section in which each threshold was allowed to vary according to its own distribution, using baseline-category and continuation-ratio logits. Applying this approach to the cumulative logit model is more difficult, however, due to the required ordering of the thresholds. Recall that for the cumulative logit model, the thresholds are ordered such that $\alpha_1 < \cdots < \alpha_q$. For the random threshold model the restriction becomes

$$\alpha_1 + u_{i1} < \cdots < \alpha_q + u_{iq},$$

where $\mathbf{u}_i = (u_{i1}, \cdots, u_{iq})'$ is assumed to be multivariate normal with mean $\mathbf{0}$ and covariance matrix Σ_α. Unless the thresholds are well separated or the diagonal elements of Σ_α are small, this ordering is likely to be violated during the estimation routine. Indeed, $P(\alpha_1 + u_{i1} < \cdots < \alpha_q + u_{iq}) < 1$ unless $\mathrm{corr}(u_{ir}, u_{ir'}) = 1$ for $r \neq r'$ (Tutz and Hennevogl 1996). There are two approaches that can be used to avoid such violations. First, one could use a constrained maximization routine which would enforce the restriction while carrying out the maximization. Such algorithms are difficult to program, especially when using quadrature or Monte Carlo methods. The second approach is to transform the thresholds to a new set of parameters which are not restricted. Maximization is then carried out using the transformed thresholds. Such an approach was utilized by Tutz and Hennevogl (1996). To incorporate the reparameterized thresholds into the multivariate generalized linear model framework, new response functions and design matrices must be defined. In the next section we present the extended random effects model and outline the necessary modifications to the fitting algorithms. We then illustrate the extended model in Section 3.7.2 using the wine dataset analyzed previously in Section 3.6.1. In Section 3.7.3 we examine the extended model in more detail and discuss some of the problems that we encountered while using it. We conclude in Section 3.7.4 with a small simulation study that examines the bias in regression parameters when one fits the simpler shifted threshold model in lieu of the extended model.

3.7.1 Extension to Random Thresholds

To avoid violation of the threshold ordering, Tutz and Hennevogl (1996) proposed the following reparameterization, for which the model with random thresholds is more appropriate:

$$\tilde{\alpha}_1 = \alpha_1, \quad \tilde{\alpha}_r = \log(\alpha_r - \alpha_{r-1}), \quad r = 2, \cdots, q, \tag{3.92}$$

or equivalently

$$\alpha_1 = \tilde{\alpha}_1, \quad \alpha_r = \alpha_1 + \sum_{i=2}^{r} \exp(\tilde{\alpha}_r), \quad r = 2, \cdots, q.$$

Under this parameterization, the new thresholds $\tilde{\alpha}_1, \cdots, \tilde{\alpha}_q$ are unrestricted with parameter space $(\tilde{\alpha}_1, \cdots, \tilde{\alpha}_q) \in \mathbb{R}^q$. Note that $\tilde{\alpha}_1$ is equivalent to α_1, while the remaining reparameterized thresholds measure the distance between the original thresholds.

Because of the reparameterized thresholds, the response functions and design matrices given in Section 2.4.3 for the cumulative logit link are no longer valid. In particular, a new response function must be defined so that the usual form of the multivariate generalized linear model is obtained. For the random threshold model the new response function has the form $\mathbf{h}(\boldsymbol{\eta}_{ij}) = (h_1(\boldsymbol{\eta}_{ij}), \cdots, h_q(\boldsymbol{\eta}_{ij}))'$ with

$$
\begin{aligned}
h_1(\boldsymbol{\eta}_{ij}) &= \frac{1}{1+\exp(-\eta_{ij1})} \\
h_r(\boldsymbol{\eta}_{ij}) &= \frac{1}{1+\exp(-\eta_{ij1} - \sum\limits_{l=2}^{r} e^{\eta_{ijl}})} - \frac{1}{1+\exp(-\eta_{ij1} - \sum\limits_{l=2}^{r-1} e^{\eta_{ijl}})}, \quad r = 2, ..., q,
\end{aligned}
\tag{3.93}
$$

where the linear predictor $\boldsymbol{\eta}_{ij} = (\eta_{ij1}, \cdots, \eta_{ijq})'$ is given by

$$
\begin{aligned}
\eta_{ij1} &= \tilde{\alpha}_1 + \mathbf{x}_{ij}^{'}\boldsymbol{\beta} + u_{i1}, \\
\eta_{ijr} &= \tilde{\alpha}_r + u_{ir}, \quad r = 2, \cdots, q.
\end{aligned}
\tag{3.94}
$$

Equivalently, the new link function $\mathbf{g}(\boldsymbol{\pi}_{ij}) = (g_1(\boldsymbol{\pi}_{ij}), \cdots, g_q(\boldsymbol{\pi}_{ij}))'$ is given by

$$
\begin{aligned}
g_1(\boldsymbol{\pi}_{ij}) &= \log\left(\frac{\pi_{ij1}}{1-\pi_{ij1}}\right), \\
g_r(\boldsymbol{\pi}_{ij}) &= \log\left[\log\left(\frac{\sum\limits_{l=1}^{r}\pi_{ijl}}{1-\sum\limits_{l=1}^{r}\pi_{ijl}}\right) - \log\left(\frac{\sum\limits_{l=1}^{r-1}\pi_{ijl}}{1-\sum\limits_{l=1}^{r-1}\pi_{ijl}}\right)\right]. \quad r = 2, ..., q.
\end{aligned}
\tag{3.95}
$$

To accommodate the new link function, the design matrix now takes the form

$$
Z_{ij} = \begin{bmatrix} 1 & & & \mathbf{x}_{ij}^{'} \\ & 1 & & 0 \\ & & \ddots & \vdots \\ & & 1 & 0 \end{bmatrix}. \tag{3.96}
$$

Using (3.95) and (3.96), the random threshold model has the form $\mathbf{g}(\boldsymbol{\pi}_{ij}) = Z_{ij}\boldsymbol{\beta} + \mathbf{u}_i$ where $\boldsymbol{\beta} = (\tilde{\alpha}_1, \cdots, \tilde{\alpha}_q, \boldsymbol{\gamma}^{'})'$, $\mathbf{u}_i = (u_{i1}, \cdots, u_{iq})'$, and $\boldsymbol{\gamma}$ is the fixed effects parameter vector. For the more general model that allows both threshold random effects and cluster-specific random effects $\mathbf{w}_{ij}$, the model takes the usual form $\mathbf{g}(\boldsymbol{\pi}_{ij}) = Z_{ij}\boldsymbol{\beta} +$

$W_{ij}\mathbf{u}_i$ where the random effects design matrix has the form

$$W_{ij} = \begin{bmatrix} 1 & & & & \mathbf{w}'_{ij} \\ & 1 & & & 0 \\ & & \ddots & & \vdots \\ & & & 1 & 0 \end{bmatrix},$$

and $\mathbf{u}_i \sim MVN(\mathbf{0}, \Sigma)$.

The algorithms given in Section 3.3 and 3.5 can be used to fit the random thresholds model by utilizing the new response function (3.93) and the design matrix (3.96). In addition, the first and second derivatives of the response function with respect to the linear predictor must also be updated. Let $\Gamma_{ij1} = \dfrac{1}{1+\exp(-\eta_{ij1})}$ and $\Gamma_{ijr} = \dfrac{1}{1+\exp(-\eta_{ij1} - \sum_{l=2}^{r} \exp(\eta_{ijl}))}$, $r = 2, \cdots, q$. Also let $c_{ijr} = \exp(\eta_{ijr})$, $r = 2, \cdots, q$, with $c_{ij1} = 1$. The first derivative matrix $D_{ij} = \dfrac{d\mathbf{h}(\boldsymbol{\eta}_{ij})}{d\boldsymbol{\eta}_{ij}}$ has a (u, v)th element of

$$\begin{aligned} d_{uv} &= 0 && \text{if } u > v, \\ &= \Gamma_{ijv}(1 - \Gamma_{ijv})c_{ijv} && \text{if } u = v, \\ &= \Gamma_{ijv}(1 - \Gamma_{ijv})c_{iju} - \Gamma_{ij,v-1}(1 - \Gamma_{ij,v-1})c_{iju} && \text{if } u < v. \end{aligned}$$

To calculate the observed information matrix, the second derivative matrix of the response function with respect to the linear predictor is required. Formulas for the second derivative matrix in the varying threshold model are very complicated. We programmed such formulas for the application considered in this section but do not provide the details. As an alternative, one can use numerical derivatives to calculate the observed information matrix.

3.7.2 Application: Wine Tasting Dataset

To illustrate the cumulative logit model with varying thresholds, Tutz and Hennevogl (1996) analyzed the wine tasting dataset (Table 3.2) using their Monte Carlo EM algorithm. Recall that judges were asked to rate the bitterness of the wine on a five-point scale. In the varying threshold model, each judge is allowed to have differing thresholds. The linear predictor $\boldsymbol{\eta}_{ij}$ for the jth rating from the ith judge has the form

$$\begin{aligned}\eta_{ij1} &= \tilde{\alpha}_1 + \beta_{TE}\ x_{ij1} + \beta_{CO}\ x_{ij2} + \beta_{BO}\ x_{ij3} + u_{i1}, \qquad (3.97)\\ \eta_{ijr} &= \tilde{\alpha}_r + u_{ir}, \quad r = 2, \cdots, 4,\end{aligned}$$

where the $\{\tilde{\alpha}_r\}$ are defined as in (3.92) and $\mathbf{u}_i' = (u_{i1}, \cdots, u_{i4})$ is assumed to be multivariate normal with mean $\mathbf{0}$ and covariance matrix Σ. The regression parameters and covariates in (3.97) are defined as in model (3.87) in Section 3.6.1. The dimension of the random effects vector $\mathbf{u}_i$ is four, thus the unstructured covariance matrix Σ contains ten parameters: four variance terms and six covariance terms. One might consider reducing the number of parameters in Σ by, for example, assuming common correlations between the random effects. However, such assumptions are typically not possible for this extended model which we will discuss in more detail in the next section.

Table 3.11 contains the regression parameter and variance component estimates for model (3.97) as reported by Tutz and Hennevogl (1996), and those obtained using our adaptive Gauss-Hermite quadrature algorithm. We have also included the results from the shifted threshold model fit in Section 3.6.1. We note, however, that the threshold estimates for the second, third, and fourth thresholds of this model are not comparable with the corresponding thresholds from the extended models. Due to the small estimated variance components and large correlations, we modified our adaptive algorithm so that the Cholesky square root of Σ was estimated. To fit model (3.97) we initially used a small number of quadrature points (5) to obtain starting values, and then refit the model using 15 quadrature points in each dimension. Thus, for each of the nine judges at each iteration, $15^4 = 50{,}625$ quadrature points were evaluated. The algorithm required approximately 58 hours to converge. We then ensured that parameter estimates had converged to four decimal places by refitting the model using 16 quadrature points in each dimension, starting from the final parameter estimates of the previous fit.

We begin by examining the results reported by Tutz and Hennevogl (1996) using the EM algorithm with 10, 20, and 30 Monte Carlo samples. First note that the estimates are quite variable across the three simulation sizes, and do not appear to be settling down. Even the log-likelihood values are quite different from the results with 10 samples (-78.100) and the results with 30 samples (-80.437). We again see extremely large estimates for the standard errors, especially with the simulation size of 30. We assume that the large estimates are a result of the error in their programming that was noted before. The large variation in the parameter estimates in most likely a result of using Monte Carlo samples sizes that are too small. In addition, Tutz and Hennevogl (1996) do not account for the Monte Carlo error in the numerical integration which would propagate through to the parameter estimates. Tutz and Hennevogl (1996) concluded, by comparison of log-likelihoods, that the inclusion of threshold specific random effects for this dataset was unnecessary.

Comparing the results of Tutz and Hennevogl (1996) to the adaptive algorithm we see a number of differences. The most dramatic difference is found in the estimate of the standard deviation for the first threshold where the adaptive algorithm obtained 1.843 while the Monte Carlo EM algorithm with 30 samples obtained 2.942. Since the first threshold is the same under both the varying threshold parameterization and the shifted threshold model, the estimated standard deviation for the first threshold of the former model is comparable with shifted threshold model. Indeed, Tutz and

Table 3.11: Parameter and variance component estimates, and log-likelihood values (LL) for fitting model (3.97) to the wine tasting dataset using the cumulative logit link. Numbers in column labels denote the number of quadrature points or Monte Carlo samples used in the adaptive Gauss-Hermite (AGH) and Monte Carlo EM (MCEM) (Tutz and Hennevogl 1996) algorithms. SHIFTED denotes the results obtained using the AGH algorithm with five quadrature points allowing only a shifting of thresholds.

	SHIFTED	AGH(15)	MCEM(10)	MCEM(20)	MCEM(30)
$\tilde{\alpha}_1$	-4.082	-4.661	-6.817	-4.230	-4.916
$\tilde{\alpha}_2$	-0.930	1.287	1.632	1.313	1.410
$\tilde{\alpha}_3$	1.797	0.995	0.951	0.860	0.945
$\tilde{\alpha}_4$	3.657	0.560	0.717	0.529	0.563
β_{TE}	1.536	1.554	1.529	1.546	1.529
	(.298)	(.302)	(.628)	(.938)	(1.144)
β_{CO}	0.916	0.922	0.943	0.976	0.971
	(.256)	(.247)	(.464)	(.744)	(.875)
β_{BO}	0.122	0.123	0.093	0.093	0.075
	(.232)	(.234)	(.424)	(.644)	(.764)
σ_1	1.145	1.843	3.206	2.775	2.942
σ_2	–	0.223	0.412	0.324	0.353
σ_3	–	0.225	0.400	0.344	0.425
σ_4	–	0.178	0.141	0.296	0.154
LL	-81.394	-80.898	-78.100	-78.895	-80.437

Hennevogl (1996) commented that the varying threshold model provided a distinctly larger estimate for this threshold standard deviation (2.942 versus 1.145). However, we see with the adaptive algorithm that the estimates are not as markedly different (1.843 versus 1.145). We do see slightly larger parameter estimates and standard errors for the adaptive algorithm when compared with the shifted model, but in general the two models are providing very similar results. Without a formal test for comparing the two models, the difference in the log-likelihoods would certainly suggest that the more complicated model with nine additional parameters is unneeded.

We mentioned before that we needed to estimate the Cholesky factor of Σ instead of Σ for this dataset. Due to the parameterization of the thresholds given in (3.92), the correlations between thresholds can be extremely high. For the given dataset, the estimated correlation matrix is given by

$$\begin{bmatrix} 1 & -.951 & -.135 & -.588 \\ & 1 & .438 & .690 \\ & & 1 & -.724 \\ & & & 1 \end{bmatrix}.$$

Such large correlations coupled with small variance component estimates can cause problems for the estimation algorithm. Use of the Cholesky form of the covariance matrix can help alleviate some of these problems.

3.7.3 Discussion

As in the varying threshold models using the baseline-category and continuation-ratio logit links, the extended threshold model of Tutz and Hennevogl (1996) provides a way of relaxing the shifted threshold assumption for the cumulative logit link. Indeed for situations like the wine tasting experiment where subjects are asked to give their preferences for items, one might expect the thresholds to vary individually across subjects. In contrast, however, to the varying threshold models for the baseline-category and continuation-ratio logit links, the extended threshold model of Tutz and Hennevogl (1996) does not have the same interpretation for the shifting of thresholds.

Consider a model with three responses and, thus, two thresholds (α_1, α_2). For the previous varying threshold models, the thresholds were allowed to vary by introducing random effects linearly with the thresholds: $(\alpha_1 + u_{i1}, \alpha_2 + u_{i2})$, where $\mathbf{u}_i' = (u_{i1}, u_{i2})$ was assumed to be multivariate normal with mean $\mathbf{0}$ and covariance Σ. Due to the ordering of thresholds in the cumulative logit model, random effects were introduced linearly with the reparameterized thresholds: $(\tilde{\alpha}_1 + u_{i1}, \tilde{\alpha}_2 + u_{i2})$, with the same assumptions on the random effects. Since $\tilde{\alpha}_1 = \alpha_1$, the shifted first threshold in the extended model is the same as that for the previous varying threshold models. For the second threshold, however, we have

$$\begin{aligned}\tilde{\alpha}_2 + u_{i2} &= \log(\alpha_2 - \tilde{\alpha}_1) + u_{i2} \\ &= \log(\alpha_2 - \tilde{\alpha}_1) + \log(\exp(u_{i2})) \\ &= \log(\alpha_2 \exp(u_{i2}) - \tilde{\alpha}_1 \exp(u_{i2})).\end{aligned}$$

Thus we lose the simplistic interpretation of shifting for the second threshold. We also see that $\tilde{\alpha}_1$ and $\tilde{\alpha}_2$ may be highly correlated as $\tilde{\alpha}_2$ contains $\tilde{\alpha}_1$. This was already seen for the wine dataset were the first two thresholds had an estimated correlation of -0.951.

As was noted for the wine tasting dataset, and from our experience with other datasets, fitting the varying threshold model can be difficult. For one, it is often difficult to find starting values for the covariance matrix Σ, especially when the number of thresholds is large. The reason is that the first variance component measures the variability in the first threshold, while the remaining measure the variability in the log difference between adjacent thresholds. Thus, the variance estimates are often quite different in magnitude. For the wine tasting dataset, the standard deviation for the first threshold was over 10 times that of the last threshold. Such disparity in the estimates can also cause numerical problems in the estimating routine. For the models in the previous sections, covariance (or correlation) terms were always set at zero as starting values. This typically can not be done for the varying threshold model, however, as the random effects are often highly correlated. Indeed, for the wine tasting dataset we were unable to even evaluate the likelihood until we had chosen appropriate covariance terms. To reduce the number of parameters and possibly ease the estimation problems, one could assume a structure for the correlations of the random thresholds. Unfortunately, the correlations often vary both in magnitude and sign across all pairs of thresholds. Thus forcing a set structure on the correlations can lead to more numerical problems. For example, we tried an autoregressive structure for the wine dataset with heterogeneous variances for the thresholds and correlations that declined exponentially with distance. That is

$$\begin{bmatrix} \sigma_1^2 & \rho & \rho^2 & \rho^3 \\ \rho & \sigma_2^2 & \rho & \rho^2 \\ \rho^2 & \rho & \sigma_3^2 & \rho \\ \rho^3 & \rho^2 & \rho & \sigma_4^2 \end{bmatrix}.$$

However, we were unable to even find suitable starting values to start the estimation routine.

3.7.4 Simulation Study

Due to the difficulties in fitting the varying threshold model, the unexpected results when the response categories are reversed, and the lack of the software for fitting the model, one would most likely, in practice, fit the simpler shifted threshold model. The shifted threshold model is attractive as it is relatively simple to fit and has a much simpler interpretation in terms of the original thresholds. What is not known, however, is how well the shifted threshold model performs when the varying threshold model truly holds. To examine its performance one should, ideally, simulate data with varying thresholds, fit the data using both the varying threshold model and the shifted threshold model, and then compare the results. Unfortunately, as discussed in the previous section, the varying threshold model can be very unstable and we were unable to consistently fit the model to the simulated data. Thus, we carried out a simulation study using only the shifted threshold model.

The goal of the simulation study was to determine the bias in the regression parameter estimates when data simulated with varying thresholds was fit using the shifted threshold model. To this end we simulated data from the following model

$$\eta_{ijr} = \alpha_r + x_{ij}\beta + u_{ir}, \tag{3.98}$$

$$i = 1, \cdots, n, \quad r = 1, \cdots, q = R - 1, \quad j = 1, \cdots, T,$$

where $R = 3$, $T = 7$, $n = 100$, and $\mathbf{u}_i' = (u_{i1}, u_{i2})$ is multivariate normal with mean $\mathbf{0}$ and covariance Σ. The covariate values $\{x_{ij}\}$ were simulated from the standard normal distribution, and the parameters in (3.98) were given the values $\alpha_1 = -1.25$, $\alpha_2 = 1.25$, and $\beta = 0.5$. We chose thresholds with a large spread between them to reduce the chance of them overlapping when perturbed by a random effect. For the simulations we varied the structure for the covariance matrix Σ of the threshold random effects $\mathbf{u}_i$. That is, we varied $(\sigma_1^2, \rho, \sigma_2^2)$ where σ_r^2 denotes the variance component for the rth threshold, $r = 1, 2$, and ρ denotes the correlation between the thresholds. For the variance components we considered two situations: an extreme situation in which the variance components were quite different and a situation where the thresholds had similar variabilities. We then varied the correlation between the random effects, allowing both positive and negative correlations. The combinations of the factor levels given in the table below provide the settings for the 15 simulations. It is also informative to consider the ideal situation in which the data truly come from the shifted threshold model. Thus we performed three simulations where both thresholds were shifted by the same random effect with variances 0.16, 0.5, and 1.0. From a pilot study it was determined that simulation sizes of 300 would achieve Monte Carlo error estimates for the regression parameter of less then 0.01.

ρ	(σ_1^2, σ_2^2)
-0.8	(0.16, 1.0)
-0.5	(1.0, 0.16)
0	(0.5, 0.7)
0.5	
0.8	

We obtained estimates for the shifted threshold models using the adaptive Gauss-Hermite algorithm with 15 quadrature points. The initial starting value for the variance component of the random intercept was 0.5, while the estimates obtained from the fixed effects version of model (3.98) were used for the remaining parameters. For each simulation we recorded the estimated bias in the parameter estimates (where the bias of $\hat{\theta}$ is defined as $E(\hat{\theta}) - \theta$), the average standard error of $\hat{\beta}$ calculated from the observed information matrix, and the Monte Carlo standard error of $\hat{\beta}$ obtained over all the simulations. We also estimated the standard deviation of the random effect for the shifted threshold.

Tables 3.12 and 3.13 contain the results of the 15 simulations, while Table 3.14 contains the results of the three ideal runs. In Table 3.12 are the results for the extreme situation where the variances (expressed as standard deviations) were very different, broken down by the value of the correlation ρ. Since the variances were quite different, we ran simulations with the large variance component associated with the first threshold and also with the second threshold. As one would expect, there was considerable bias in the estimated threshold that coincided with the larger variance component, regardless of the correlation. We do see, however, that the largest biases for the thresholds occurred when the correlation was negative. For example, when the correlation was -0.8, the second threshold with $(\sigma_1, \rho, \sigma_2) = (.4, -0.8, 1.0)$ had an estimated bias of -0.165, or a percent bias of $\frac{\text{Bias}}{\text{True Value}} = \frac{-0.165}{1.25} = -13.2\%$. The threshold corresponding to the smaller variance component had smaller estimated bias, ranging from -0.090 to 0.075.

The shifted threshold model provided biased estimates for the regression parameter β as well. The largest estimated bias occurred when $(\sigma_1, \rho, \sigma_2) = (1.0, -0.8, .4)$ with a value of -0.029, or a percent bias of -5.8%. In fact there is a clear pattern seen in Table 3.12 that larger estimated biases occurred when the correlation was zero or negative. For positive correlations, the largest bias was -0.014 (percent bias of -2.8%). We also see that, for this particular model, larger biases generally occurred in β when the larger variance component was associated with the first threshold. An exception occurred for the $\rho = 0$ case. These differences, however, could be due to Monte Carlo error which was less than 0.01.

Table 3.12 also contains the average estimate of the standard deviation for the shifted threshold σ_A. This estimate is not very meaningful as the true thresholds vary individually. However, it is interesting to note that the smallest estimates of σ_A were found in the $\rho = -0.8$ case and were about 0.28 while the remaining simulations produced averages between 0.45 and 0.7. We also included the estimates of the standard error for β from the observed information matrix and the Monte Carlo estimate. Neither are very informative though. Ideally we should compare the observed information estimate to that obtained in the varying threshold model to see if they agree. But due to the instability of the varying threshold model, we were not able to accomplish this.

We now examine the second table, Table 3.13, which contains the results using similar variance component estimates. Since the variance component estimates were so similar we only ran the simulations for the given ordering of the variance components. We see similar patterns for these results as seen in Table 3.12. Namely, larger estimated biases occurred in the threshold parameters and the regression coefficient β when the correlation was zero or negative. Indeed, when the correlation was positive

Table 3.12: Estimated bias of parameter estimates for the shifted threshold model using data simulated from the varying threshold model with extreme variation differences in the thresholds. $\mathrm{SE}(\hat{\beta})_O$ and $\mathrm{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations. The estimate σ_A is the average standard deviation estimate of the shifted thresholds.

	THRESHOLD COVARIANCE STRUCTURE			
	$(\sigma_1, \rho, \sigma_2)$			
	(.4, ρ, 1)	(1, ρ, .4)	(.4, ρ, 1)	(1, ρ, .4)
	$\rho = 0$			
α_1	-0.050	0.132		
α_2	-0.142	0.017		
β	-0.023	-0.018		
$(\mathrm{SE}(\hat{\beta})_O)$	(.075)	(.077)		
$(\mathrm{SE}(\hat{\beta})_{MC})$	(.073)	(.072)		
σ_A	0.559	0.554		
	$\rho = .5$		$\rho = -.5$	
α_1	-0.068	0.116	-0.039	0.144
α_2	-0.123	0.049	-0.155	0.008
β	-0.007	-0.011	-0.021	-0.023
$(\mathrm{SE}(\hat{\beta})_O)$	(.076)	(.079)	(.075)	(.077)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(.079)	(.076)	(.078)	(.075)
σ_A	0.633	0.620	0.484	0.633
	$\rho = .8$		$\rho = -.8$	
α_1	-0.090	0.108	-0.024	0.152
α_2	-0.117	0.075	-0.165	-0.001
β	0.004	-0.014	-0.020	-0.029
$(\mathrm{SE}(\hat{\beta})_O)$	(.076)	(.079)	(.075)	(.075)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(.076)	(.077)	(.070)	(.074)
σ_A	0.685	0.670	0.274	0.275

Table 3.13: Estimated bias of parameter estimates for the shifted threshold model using data simulated from the varying threshold model with similar variation differences in the thresholds. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations. The estimate σ_A is the average standard deviation estimate of the shifted thresholds.

	THRESHOLD COVARIANCE STRUCTURE $(\sigma_1, \rho, \sigma_2)$	
	(.71, ρ, .84)	(.71, ρ, .84)
	$\rho = 0$	
α_1	0.031	
α_2	-0.087	
β	-0.026	
(SE($\hat{\beta}$)$_O$)	(.073)	
(SE($\hat{\beta}$)$_{MC}$)	(.074)	
σ_A	0.557	
	$\rho = .5$	$\rho = -.5$
α_1	0.006	0.054
α_2	-0.059	-1.113
β	-0.003	-0.027
(SE($\hat{\beta}$)$_O$)	(.073)	(.075)
(SE($\hat{\beta}$)$_{MC}$)	(.074)	(.074)
σ_A	0.669	0.411
	$\rho = .8$	$\rho = -.8$
α_1	-0.022	0.051
α_2	-0.038	-0.127
β	0.003	-0.034
(SE($\hat{\beta}$)$_O$)	(.078)	(.074)
(SE($\hat{\beta}$)$_{MC}$)	(.080)	(.071)
σ_A	0.736	0.326

Table 3.14: Estimated bias of parameter estimates for the shifted threshold model using data simulated from the shifted threshold model. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations. The estimate σ_A is the average standard deviation estimate of the shifted thresholds.

	SHIFTED THRESHOLD STANDARD DEVIATION		
	$\sigma = .4$	$\sigma = .71$	$\sigma = 1$
α_1	-0.009	-0.018	-0.019
α_2	-0.005	-0.007	-0.013
β	-0.002	-0.002	-0.002
(SE($\hat{\beta}$)$_O$)	(.076)	(.080)	(.081)
(SE($\hat{\beta}$)$_{MC}$)	(.077)	(.083)	(.080)
σ_A	0.394	0.706	1.000

the estimated bias for the regression coefficient β was only (in absolute value) 0.003, or a percent bias of less than one percent. The largest bias in the regression coefficient (-0.034) was found with $\rho = -0.8$. It is also interesting to note that the estimated standard deviation for the random effect increased from 0.326 with $\rho = -0.8$ to 0.736 with $\rho = 0.8$. This latter value being near the average (0.775) of the two standard deviation values assigned to the varying thresholds.

In Table 3.14 are the results when the data truly come from the shifted threshold model. As one would expect, the shifted threshold model fit the true data accurately. In this ideal situation, the absolute estimated bias of the regression parameter for all three shifted threshold variabilities was less than 0.01, which can be explained by the Monte Carlo error. Note that these estimated biases are very similar to those found in Table 3.13 for the similar threshold variances with positive correlation. However, the estimated bias for the threshold parameters is much less under the true model. The estimated standard deviations of the thresholds are very close to the true values. In addition, the standard errors from the observed information matrix and the Monte Carlo estimate are very similar.

From these simulations, we see that the bias in the regression parameter, the parameter of interest, decreases as the correlation between the threshold random effect increases and the variance components for the thresholds become more similar. In fact, the estimated bias of the regression parameter with similar threshold variances and positive correlation is similar to that obtained when the data is simulated under the true model. The reason for this is made clear by considering the following. In simulating the values of the random effects for the varying threshold model, we used a bivariate normal distribution with covariance matrix

$$\begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}. \tag{3.99}$$

The covariance matrix for the shifted threshold model can be viewed in the same form as (3.99), but with $\rho = 1$ and $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Thus as ρ approaches positive 1.0, and the variance components for the varying threshold become more similar, the shifted threshold model should perform well.

It is difficult to say with certainty how thresholds vary in "real life" datasets when the data are collected longitudinally or in clusters. Recall that in the latent variable motivation for the cumulative logit model (Section 3.2.2), the thresholds partitioned the underlying continuous response Y^* into a coarser, categorized version, Y. That is

$$Y = r \quad \Leftrightarrow \quad \alpha_{r-1} < Y^* \leq \alpha_r, \quad r = 1, \cdots, R,$$

with $\alpha_0 = -\infty$ and $\alpha_R = \infty$. The shifted threshold model moves the "windows" (distance between thresholds) for each response along the underlying continuous response scale, but keeps them the same size. The varying threshold model allows the "windows" to move and change size. This would allow, for example, some subjects to have wider "windows" for particular responses, and smaller for others. We feel that if the thresholds did vary individually, they would probably move in similar directions (i.e. be positively correlated) for which the shifted threshold model would provide reasonable estimates.

An open question is how the standard errors relate between the two models. We would expect that the standard error estimates from the two models would be similar, especially as the variance components for the thresholds become similar and the correlation nears 1.0. From the results of the wine tasting dataset it was seen that the standard errors between the more complex model and the simpler model were quite similar. Ideally, one would like to account for all sources of extraneous variability when modeling. However, when doing so produces a much more complicated model for both fitting and interpretation, use of a simpler model that attempts to account for most of the heterogeneity is probably more appropriate.

CHAPTER 4
NONPARAMETRIC MAXIMUM LIKELIHOOD ESTIMATION IN MULTIVARIATE GENERALIZED LINEAR MIXED MODELS

4.1 Introduction

In the previous chapter, we considered random effects models for nominal and ordinal response data where we assumed that the random effects followed a multivariate normal distribution. The multivariate normal assumption allows for a variety of covariance structures for the modeling of the random effects. Another advantage of such an assumption is that the random effects distribution can be estimated more accurately when the cluster sizes are small. The motivation of generalized linear mixed models as extensions of linear mixed models has also contributed to the popularity of the normality assumption. In spite of its popularity and attractive features, the assumption of normality can rarely be verified. Thus a concern of making such an assumption is the possible misspecification of the random effects distribution.

There have been a number of studies that have investigated the effects of misspecification of the random effects distribution, one of which was the seminal paper by Heckman and Singer (1984). Examining models for censored longitudinal economic data, Heckman and Singer (1984) showed that the fixed parameter estimates in a particular Weibull regression model were highly sensitive to the parametric assumption about the random effects distribution. Though the severe changes reported in the paper were specific to the model and data combination, the results provided evidence that misspecification could impact the estimates, and ultimately the inference in a model. Others have shown supportive, although less dramatic evidence for other types of models. Davies (1987) performed a small study in which he analyzed simulated residential mobility data using a probit model. For each household, a response profile consisting of zeros and ones was simulated according to whether or not the household moved in that year. The probability of moving was dependent on five household specific covariates whose parameter values were determined from a previous study of American households. Davies (1987) simulated the data using both a normal mixing distribution and a triangular mixing distribution. To maximize the log-likelihood, Davies (1987) assumed that the mixing distribution was a continuous rectangular distribution and used Gauss-Legendre quadrature to integrate over the mixing distribution. One reason for choosing the rectangular distribution was that the numerical integration was restricted to a finite range. Davies (1987) concluded that the discrepancies in the parameter estimates are less extreme if the true and assumed mixing distributions are at least somewhat similar. More recently, Neuhaus et al. (1992) and Butler and Louis (1992) showed for the binary random intercept logistic model that the model parameters were indeed inconsistent if the random effects distribution was misspecified. The magnitude of the bias, however, was found to be typically small. Neuhaus et al. (1992) concluded that the assumption of normality resulted in generally robust inferences about the regression parameters under misspecification of the random effects distribution. In light of these conclusions, there has been a considerable amount of recent work focused on nonparametric approaches to fitting these models that avoids parametric assumptions about the distribution of the random effects (Heckman and Singer 1984; Davies 1987; Wood and Hinde 1987; Follmann and Lambert 1989; Butler and Louis 1992; Wedel and DeSarbo 1995; Aitkin 1996, 1999).

In this chapter, we propose a class of alternative models for repeated nominal and ordinal response data in which the random effects or mixing distribution is estimated nonparametrically. The proposed models, which can be considered as extensions of those by Aitkin (1996, 1999), are estimated by way of maximum likelihood estimation, providing nonparametric maximum likelihood (NPML) estimates of both the fixed parameters and the mixing distribution. We begin in Section 4.2 with the definition of the proposed models. Within this section we will describe an EM algorithm for the NPML estimation of the unknown parameters and the mixing distribution. We will discuss methods of inference as well as provide a method for calculating standard errors. In Section 4.3 we will provide sufficient conditions under which the proposed models are identifiable. We will then, in Section 4.4, apply the proposed models to the wine tasting dataset considered in the previous chapter. We conclude this chapter by reporting the results of two simulation studies. The first study compares the models of the previous chapter to the proposed models under a variety of mixing distributions. In the second study we investigate the validity of the Wald and likelihood-ratio tests within the context of the NPML model for making inference on the fixed effects.

4.2 Nonparametric Maximum Likelihood Estimation

As in Chapter 3, the proposed models will be developed as extensions of multivariate generalized linear models. The model definition and subsequent estimation algorithm will be presented for a general link function $\mathbf{g}(\boldsymbol{\pi})$ or response function $\mathbf{h}(\boldsymbol{\eta})$. Thus the form of the design matrix Z and parameter vector $\boldsymbol{\beta}$ are assumed to be appropriate for the desired link function with the specific forms being found in Chapter 2. In the discussion that follows, only models in which thresholds of the nominal or ordinal regression model are allowed to vary across subjects will be considered. In Chapter 5 we will look at extending the models to allow for bivariate random effect structures.

As before, denote the multinomial response vector for the jth observation from the ith cluster as $\mathbf{y}_{ij}$ with corresponding multinomial sample size n_{ij} and covariate vector $\mathbf{x}_{ij}$, $j = 1, \cdots, T_i$, $i = 1, \cdots, n$. We assume that conditional on an unobserved random variable u_i, $\bar{\mathbf{y}}_{ij}$ has the multivariate exponential form with response function $\mathbf{h}(\boldsymbol{\eta}_{ij})$ and linear predictor $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + u_i$. In addition we assume that observations between clusters are independent, and observations within clusters are conditionally independent. That is, we assume for the complete response vector $\mathbf{y}$ and random effects vector $\mathbf{u}$ that

$$f(\mathbf{y} \mid \boldsymbol{\beta}; \mathbf{u}) = \prod_{i=1}^{n} \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; u_i),$$

where $f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; u_i)$ denotes the multinomial distribution. Denoting the mixture distribution by G, the likelihood for the model is given by

$$L(\boldsymbol{\beta}, G) = \prod_{i=1}^{n} \int \cdots \int \left[\prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; u_i) \right] dG(u_i). \tag{4.1}$$

In Chapter 3 we estimated the parameters $\boldsymbol{\beta}$ and the mixing distribution G in (4.1) by assuming G was normal and maximizing the marginal likelihood obtained by numerical integration. To avoid possible misspecification, we now consider estimating G nonparametrically. Specifically, we assume that G is a discrete distribution function with unknown finite support size K, masses $\mathbf{p}' = (p_1, \cdots, p_K)$, and mass points $\mathbf{m}' = (m_1, \cdots, m_K)$, where $\sum_{k=1}^{K} p_k = 1$ and $p_k \geq 0,\ k = 1, \cdots, K$. Considering only the class of discrete distributions with finite support size is not restrictive, since it has been well established that the NPML estimate of a (possibly continuous) mixing distribution is concentrated on a finite number of points (Kiefer and Wolfowitz 1972; Laird 1978; Lindsay 1983a, 1983b). For fixed support size, the resulting likelihood is given by

$$L(\boldsymbol{\beta}, \mathbf{p}, \mathbf{m}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k\, f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, m_k), \tag{4.2}$$

where $f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, m_k) = \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}, m_k)$.

An abundance of literature exists in the general area of finite mixture modeling, with a thorough introduction being found in Titterington et al. (1985). The application of mixture models to the regression setting has become increasingly popular as well. There have been a number of authors who have proposed special cases of the models defined by likelihood (4.2). Follmann and Lambert (1989) considered a binary logistic regression model in which the intercept was assumed to vary according to a discrete distribution. The models that they considered could account for overdispersion at the binomial level. Lindsay et al. (1991) later related the binary logistic regression mixture model to the binary Rasch model (Rasch 1961). Butler and Louis (1992) also considered a binary logistic regression model with a random intercept, but allowed for longitudinal observations as well. We noted in Chapter 3 that Adams and Wilson (1996) and Adams et al. (1997) both assumed a discrete distribution for their shifted and varying threshold Rasch models. In their approach, however, they assumed that the mass points were known and only estimated the masses. Both Wedel and DeSarbo (1995) and Aitkin (1996, 1999) have considered nonparametric mixture extensions in generalized linear models. Aitkin (1996, 1999) generalized work by Wood and Hinde (1987) by modeling overdispersion and heterogeneity in the class of generalized linear models through the use of nonparametric mixtures. Using a similar motivation, Wedel and DeSarbo (1995) proposed a general latent class model which allowed for a generalized linear model within each class. Both Wedel and DeSarbo (1995) and Aitkin (1996, 1999) used identical EM algorithms for estimation of the parameters, though Wedel and DeSarbo (1995) estimated a set of parameters for each latent class. We now present a similar EM algorithm for maximizing (4.2). Following the section on estimation, we will discuss in Section 4.2.2 methods for making inference in the NPML approach.

4.2.1 Estimation

In likelihood (4.2) we have replaced the intractable normal integrals from the previous chapter with finite summations. Though this would seem to be a great simplification, finding the maximum likelihood values for $\boldsymbol{\beta}$, $\mathbf{p}$, $\mathbf{m}$, and K is still problematic and often computationally intensive. Of most difficulty is determining the location of the mass points $\mathbf{m}$. The approaches for maximizing (4.2) can be generally grouped into those that treat K as fixed and those that estimate K within the maximization algorithm. The EM algorithm that we propose belongs to the former group, however we will first briefly discuss some of the algorithms in the latter. A complete discussion of both types of algorithms can be found in Bohning (1995).

Many of the algorithms that estimate K within the maximization procedure are based on the geometric interpretation of mixture models as described by Lindsay (1983a, 1983b). Lindsay (1983a) showed that finding the maximum likelihood estimate of a mixture distribution was equivalent to maximizing a concave function over a convex set. Thus methods used for convex optimization can be applied to mixture models as well. In particular, directional derivatives or gradient functions can be used as directional guides for finding the maximum. Such functions are utilized in a number of vertex direction algorithms such as the vertex direction method (Lindsay 1983a), the vertex exchange method (Lesperance and Kalbfleisch 1992), and the intra simplex direction method (ISDM) (Lesperance and Kalbfleisch 1992). Such methods either sequentially increase the support size at each iteration or adaptively add or subtract points at each iteration. Follmann and Lambert (1989) utilized a combination of directional derivatives and the EM algorithm for maximizing mixtures of binomials. In general, methods that estimate the support size adaptively are complex to program. However, Lesperance and Kalbfleisch (1992) have shown that convergence for certain variants, such as the ISDM, can be extremely fast.

An alternative method for finding the maximum likelihood estimate of a mixing distribution is to treat K as fixed. The method is based on a result mentioned previously that the nonparametric maximum likelihood estimate has a finite support size. Using this result, one can sequentially maximize the likelihood for fixed $K = 1, 2, \cdots$ until convergence is obtained. The EM algorithm is a popular algorithm for computing the maximum at each fixed K because of its numerical simplicity and guaranteed monotonicity. However it can be extremely slow and convergence to a local maximum is possible. In light of this, we first outline the basic EM algorithm for maximizing (4.2). We then discuss methods for accelerating the convergence of the algorithm. We conclude this section with comments concerning the application of the algorithm.

EM algorithm

We use the EM algorithm to estimate the regression parameters $\boldsymbol{\beta}$, the mass points $\mathbf{m}$, and the mass $\mathbf{p}$. The implementation of the EM algorithm relies on the complete log-likelihood of both the observed and unobserved data. For the given models, the complete log-likelihood has the form

$$\log L(\boldsymbol{\beta}, G) = \sum_{i=1}^{n} \log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, u_i) + \sum_{i=1}^{n} \log G(u_i), \tag{4.3}$$

where $G(u_i)$ is the K component discrete distribution function for u_i with mass points $\mathbf{m}$ and masses $\mathbf{p}$. The complete log-likelihood is maximized using an iterative EM algorithm. In the E-step the complete log-likelihood is replaced by its expectation, calculated on the basis of provisional estimates of $\boldsymbol{\beta}$ and G. This expectation is then maximized in the M-step with respect to $\boldsymbol{\beta}$ and G to obtain new provisional estimates. These two steps are then alternated until no further improvement in the likelihood occurs. The details for the E-step and M-step now follow.

In the E-step at the $(s+1)$th iteration, the expectation of the complete log-likelihood (4.3) is calculated with respect to the conditional distribution $f(\mathbf{u} \mid \mathbf{y}, \boldsymbol{\beta}^{(s)}, \mathbf{m}^{(s)}, \mathbf{p}^{(s)})$ where $\boldsymbol{\beta}^{(s)}, \mathbf{m}^{(s)}$, and $\mathbf{p}^{(s)}$ are the estimated parameters from the previous iteration. Using independence, Bayes Rule, and expressing $G(u_i)$ in terms of the masses $\mathbf{p}$ and mass points $\mathbf{m}$, one obtains the expectation

$$E[\log L(\boldsymbol{\beta}, \mathbf{m}, \mathbf{p} \mid \boldsymbol{\beta}^{(s)}, \mathbf{m}^{(s)}, \mathbf{p}^{(s)})] =$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \frac{[\log f(\tilde{\mathbf{y}}_i \mid m_k, \boldsymbol{\beta}) + \log p_k] \;\; p_k^{(s)} f(\tilde{\mathbf{y}}_i \mid m_k^{(s)}, \boldsymbol{\beta}^{(s)})}{\sum\limits_{l=1}^{K} p_l^{(s)} f(\tilde{\mathbf{y}}_i \mid m_l^{(s)}, \boldsymbol{\beta}^{(s)})}$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} [\alpha_{ik}^{(s)} \log f(\tilde{\mathbf{y}}_i \mid m_k, \boldsymbol{\beta}) + \alpha_{ik}^{(s)} \log p_k], \tag{4.4}$$

where

$$\alpha_{ik}^{(s)} = \frac{p_k^{(s)} f(\tilde{\mathbf{y}}_i \mid m_k^{(s)}, \boldsymbol{\beta}^{(s)})}{\sum\limits_{l=1}^{K} p_l^{(s)} f(\tilde{\mathbf{y}}_i \mid m_l^{(s)}, \boldsymbol{\beta}^{(s)})}.$$

The $\{\alpha_{ik}^{(s)}\}$ can be interpreted as the estimated posterior probability that the response vector $(\mathbf{y}_{i1}^{'}, ..., \mathbf{y}_{iT_i}^{'})$ for subject i comes from component k. Also note that the $\{\alpha_{ik}^{(s)}\}$ are known constants depending only on the parameter estimates from the previous iteration.

The M-step consists of maximizing (4.4), the expectation of the complete log-likelihood, with respect to $\boldsymbol{\beta}$, $\mathbf{m}$, and $\mathbf{p}$. The second term on the right of (4.4) is not a function of $\boldsymbol{\beta}$ or $\mathbf{m}$ and can be maximized separately from the first term. Maximizing $\sum_i \sum_k \alpha_{ik}^{(s)} \log p_k$ subject to $\sum_{k=1}^{K} p_k = 1$ yields simply

$$\hat{p}_k^{(s)} = \sum_{i=1}^{n} \alpha_{ik}^{(s)} / n.$$

Since the $\{\alpha_{ik}^{(s)}\}$ are known, the first term on the right of (4.4),

$$\sum_{i=1}^{n}\sum_{k=1}^{K} \alpha_{ik}^{(s)} \log f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, m_k) = \sum_{i=1}^{n}\sum_{j=1}^{T_i}\sum_{k=1}^{K} \alpha_{ik}^{(s)} \log f(\tilde{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}, m_k),$$

can be recognized as the log-likelihood of a weighted multivariate generalized linear model with known weights $\{\alpha_{ik}^{(s)}\}$. Wood and Hinde (1987) noted that m_k, $k =$

$1, \ldots, K$, can be easily estimated by incorporating a K level factor in the model in place of m_k. Let $\mathbf{d}'_{ijk} = (d_{ijk1}, \ldots, d_{ijk,K-1})$ be the corresponding vector of $K-1$ dummy variables for the K level factor where

$$d_{ijkl} = \begin{cases} 1 & \text{if } l = k, \quad l = 1, \ldots, K-1 \\ 0 & \text{otherwise,} \end{cases}$$

and let $\mathcal{D}_{ijk} = \mathbf{1}_q \otimes \mathbf{d}'_{ijk}$. The linear predictor can now be re-written as $\boldsymbol{\eta}_{ijk} = Z^*_{ijk}\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^{*'} = (\boldsymbol{\beta}', m_1, \ldots, m_{K-1})$ and $Z^*_{ijk} = [Z_{ij} \mid \mathcal{D}_{ijk}]$. This new representation also requires the replication of the response vectors such that $\bar{\mathbf{y}}^*_{ijk} = \bar{\mathbf{y}}_{ij}$, $k = 1, \ldots, K$. Thus the first term on the right of (4.4) is a weighted multivariate generalized linear model with form

$$E(\bar{\mathbf{y}}^*_{ijk}) = \mathbf{h}(Z^*_{ijk}\boldsymbol{\beta}^*)$$

and known weights $\{\alpha^{(s)}_{ik}\}$.

The Fisher scoring algorithm is used to find the MLE of $\boldsymbol{\beta}^*$. The form of the score function and expected information matrix for weighted multivariate generalized linear models are known (Fahrmeir and Tutz 1994) and are given below. The score function for the algorithm is given by

$$s(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{*(s)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha^{(s)}_{ik} \sum_{j=1}^{T_i} Z^{*'}_{ijk} D_{ijk} R^{-1}_{\boldsymbol{\pi}_{ijk}} (\bar{\mathbf{y}}^*_{ijk} - \boldsymbol{\mu}_{ijk})$$

where $D_{ijk} = dh(Z^*_{ijk}\boldsymbol{\beta}^*)/d\boldsymbol{\eta}$ and $R_{\boldsymbol{\pi}_{ijk}} = \text{diag}(\boldsymbol{\mu}_{ijk}) - \boldsymbol{\mu}_{ijk}\boldsymbol{\mu}'_{ijk}$. The expected information matrix for the weighted multivariate generalize linear model is

$$F_E(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{*(s)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha^{(s)}_{ik} \sum_{j=1}^{T_i} Z^{*'}_{ijk} D_{ijk} R^{-1}_{\boldsymbol{\pi}_{ijk}} D'_{ijk} Z^*_{ijk}.$$

Thus within the M-step of the EM algorithm, the Fisher scoring algorithm is

$$\boldsymbol{\beta}^*_{p+1} = \boldsymbol{\beta}^*_p + F^{-1}_E(\boldsymbol{\beta}^*_p \mid \boldsymbol{\beta}^{*(s)})\ s(\boldsymbol{\beta}^*_p \mid \boldsymbol{\beta}^{*(s)}),$$

given starting value $\boldsymbol{\beta}_p^*$.

The NPML version of the EM algorithm with a Fisher scoring algorithm embedded in each M-step can be summarized as follows:

0. Calculate initial values $\boldsymbol{\beta}^{*(0)}$ and $\mathbf{p}^{(0)}$.

For s=0,1,2,...

1. Calculate posterior probabilities $\alpha_{ik}^{(s)}$, $i = 1, ..., n$, $k = 1, ..., K$, using $\boldsymbol{\beta}^{*(s)}$ and $\mathbf{p}^{(s)}$. Calculate $\mathbf{p}^{(s+1)}$ using $\alpha_{ik}^{(s)}$, $i = 1, ..., n$, $k = 1, ..., K$.
2. Carry out the Fisher scoring algorithm to obtain $\boldsymbol{\beta}^{*(s+1)}$ using the weights $\alpha_{ik}^{(s)}$, $i = 1, ..., n$, $k = 1, ..., K$.

The algorithm is defined in terms of a fixed support size K. To determine K, the algorithm is successively applied while incrementing K until convergence in parameter estimates and the likelihood value is obtained. We will discuss the issue of convergence as well as how one obtains starting values following the next section on EM acceleration.

Acceleration of the EM algorithm

The EM algorithm is a powerful tool for maximizing likelihoods with unobserved data and has many attractive characteristics such as guaranteed monotonicity of the log-likelihood and simplicity of implementation. One major drawback, however, is its speed of convergence which depends on the relative size of the unobserved information on the unknown parameters. In terms of mixtures, the rate of convergence will depend on the amount of information about the mixing distribution that is available from the observed data alone. The rate of convergence is also adversely affected when the unknown parameters are near the boundary of the parameter space. This can occur, for example, if one of the mass points is located at plus or minus infinity. There have been many suggestions for speeding up the convergence of the EM algorithm. We discuss two of these methods which can be used to accelerate the NPML algorithm. In the first method, the M-step of the EM algorithm is modified while maintaining the overall structure of the algorithm. In the second method, the EM algorithm is initially used and then replaced with a faster converging algorithm. Though discussed as two separate methods, one can easily combine the methods for faster convergence yet.

The first method that we will discuss, called the Expectation Conditional Maximization Either (ECME) algorithm (Liu and Rubin 1994), is an extension of the Expectation Conditional Maximization (ECM) algorithm (Meng and Rubin 1993) which itself is a generalization of the EM algorithm. To motivate the ECME algorithm we begin by outlining the EM algorithm and its extension to the ECM algorithm. Following Meng and Rubin (1993), let $\mathbf{Y} = (\mathbf{Y}_{obs}, \mathbf{Y}_{mis})$ be the complete and missing data with density $f(\mathbf{Y} \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \Theta$ is an unknown parameter vector. Also let $h(\mathbf{Y}_{obs} \mid \boldsymbol{\theta})$ denote the density of the observed data and $k(\mathbf{Y} \mid \mathbf{Y}_{obs}, \boldsymbol{\theta})$ the conditional density of $\mathbf{Y}$ given the observed data. Note that

$$h(\mathbf{Y}_{obs} \mid \boldsymbol{\theta}) \propto \int f(\mathbf{Y} \mid \boldsymbol{\theta})\, d\,\mathbf{Y}_{mis}.$$

Our goal is to find the maximum likelihood estimate, $\hat{\boldsymbol{\theta}}$, of $\boldsymbol{\theta}$ which maximizes the observed data likelihood, which we denote as

$$L(\boldsymbol{\theta}) \equiv \log h(\mathbf{Y}_{obs} \mid \boldsymbol{\theta}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') - H(\boldsymbol{\theta} \mid \boldsymbol{\theta}'), \tag{4.5}$$

where

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = E\{\log f(\mathbf{Y} \mid \boldsymbol{\theta}) \mid \mathbf{Y}_{obs}, \boldsymbol{\theta}'\}$$

is the expected complete data likelihood, and

$$H(\boldsymbol{\theta} \mid \boldsymbol{\theta}') = E\{\log k(\mathbf{Y} \mid \mathbf{Y}_{obs}, \boldsymbol{\theta}) \mid \mathbf{Y}_{obs}, \boldsymbol{\theta}'\}$$

is the expected missing data likelihood. Starting with an initial value $\boldsymbol{\theta}^{(0)}$, the EM algorithm maximizes (4.5) by iteratively maximizing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')$ over $\boldsymbol{\theta}$. That is, given the estimate $\boldsymbol{\theta}^{(s)}$ at the sth iteration, $\boldsymbol{\theta}^{(s+1)}$ is found by first computing $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)})$ as a function of $\boldsymbol{\theta}$, and then finding $\boldsymbol{\theta}^{(s+1)}$ that maximizes $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)})$. Iteration continues between the expectation step (E-step) and the maximization step (M-step) until convergence.

For some applications of the EM algorithm, the M-step can not be computed in closed form. For situations such as this, Meng and Rubin (1993) proposed the ECM algorithm which replaces the maximization of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(s)})$ with several, hopefully simpler conditional maximizations. In the ECM algorithm, the M-step is replaced by $t = 1, \cdots, T$ constrained or conditional maximization (CM) steps where in each some subset of the parameters is fixed. The CM-steps may or may not have closed forms, but since they are maximized over smaller dimensions, they are often simpler and require less time to maximize when iteration is required. Liu and Rubin (1994) proposed a generalization of the ECM algorithm called the ECM Either algorithm. In the ECME algorithm, certain CM-steps are modified such that the expectation is taken over the constrained observed data likelihood instead of the constrained expected complete data likelihood. Hence the name ECM Either arises from alternating between either the complete or observed data likelihoods within the CM-steps. Liu and Sun (1997) considered the implementation of the ECME algorithm to accelerate the EM algorithm for mixture distributions. We now modify their approach to accelerate the NPML EM algorithm discussed in the previous section.

The ECME version of the NPML EM algorithm consists of a single E-step and two CM-steps. The single E-step is the same as that for the original EM algorithm, in which we calculate the expectation defined in (4.4). The M-step, however, is separated into two CM-steps. In the first CM-step we update $\boldsymbol{\beta}^*$ by maximizing the expected conditional complete data likelihood. Hence the first CM-step corresponds to the same maximization over $\boldsymbol{\beta}^*$ that was performed in the original M-step. In the second CM-step we update $\mathbf{p}$, this time by maximizing the observed data likelihood with $\boldsymbol{\beta}^*$ fixed at its current estimate. Thus, instead of using

$$\hat{p}_k^{(s)} = \sum_{i=1}^{n} \alpha_{ik}^{(s)} / n.$$

to update $\mathbf{p}$, we update $\mathbf{p}$ by maximizing the constrained observed log-likelihood

$$L_{Obs}(\mathbf{p}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} p_k \, f(\tilde{\mathbf{y}}_i \mid \hat{\boldsymbol{\beta}}^*). \tag{4.6}$$

There are many optimizations methods that could be used to maximize (4.6). Since the first and second derivatives of L_{Obs} with respect to $\mathbf{p}$ are relatively simple to calculate, we used a Newton-Raphson algorithm for maximizing (4.6). Given a current estimate $\hat{\mathbf{p}}^{(0)}$, a new estimate is obtained as follows:

$$\hat{\mathbf{p}}^{(new)} = \hat{\mathbf{p}}^{(0)} - \delta \, \mathcal{H}^{-1} \, \mathbf{g}, \tag{4.7}$$

where $0 < \delta \leq 1$ and $\mathbf{g}$ and $\mathcal{H}$ denote the gradient vector and Hessian matrix, respectively. Let $f_{ik} = \prod_{j=1}^{T_i} f(\tilde{\mathbf{y}}_{ijk} \mid \boldsymbol{\beta}^*)$, where the product is over the T_i observations in the ith cluster with the kth dummy variable corresponding to mass point k equal to one. For the observed log-likelihood (4.6), the gradient vector and Hessian matrix have the following sth and (r, s)th elements:

$$\mathbf{g}_s = \frac{d \, L_{Obs}(p_1, \cdots, p_{K-1})}{dp_s} = \sum_{i=1}^{n} \frac{f_{is} - f_{iK}}{\sum\limits_{k=1}^{K} p_k \, f_{ik}},$$

$$\mathcal{H}_{rs} = \frac{d^2 \, L_{Obs}(p_1, \cdots, p_{K-1})}{dp_r \, dp_s} = -\sum_{i=1}^{n} \frac{(f_{ir} - f_{iK})(f_{is} - f_{iK})}{(\sum\limits_{k=1}^{K} p_k \, f_{ik})^2},$$

for $1 \leq r, s \leq K - 1$. Since the parameters $\mathbf{p}$ in (4.6) are constrained to be between zero and 1.0, the Newton-Raphson algorithm (4.7) may overshoot the feasible region of allowable parameter points. The step scaling factor δ in (4.6) can be decreased from the full step ($\delta = 1$) to partial steps ($\delta < 1$) when such violations occur. In our experience, the ECME algorithm by itself can speed convergence dramatically over the basic EM algorithm. This is especially true when the mass points are not well separated.

In a comprehensive summary of the EM algorithm within the context of mixtures, Redner and Walker (1984) noted that further research was needed in developing algorithms "...which first take advantage of the good global convergence properties of the EM algorithm by using it initially and then exploits the rapid local convergence of Newton's method or one of its variants by switching to such a method later." Indeed, even when the EM algorithm is very near the final estimates, convergence can be extremely slow due to its linear convergence rate. Thus, the second approach we mention for speeding the convergence rate of the mixture EM algorithm starts with the EM algorithm but then switches to a faster converging algorithm. Such an approach has been utilized by Follmann and Lambert (1989) for fitting mixtures of logistic regression models, and by Aitkin and Aitkin (1995) within the context of mixtures of normals. We also use this approach for the simulation studies in Section 4.5.

Any number of algorithms can be used in conjunction with the EM algorithm. We utilize the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm which was described in Section 3.3.1. Recall that it has the same structure as (4.7), but uses approximations for the gradient vector and Hessian matrix so that one does not need to program the first and second derivatives of the likelihood function. There are no specific rules as to when one should stop the EM algorithm and switch to the BFGS algorithm. Follmann and Lambert (1989) switched algorithms when two successive parameter iterates were close. For our simulation studies, we found that switching algorithms when the difference between two successive log-likelihood values was less than 0.05 had no adverse effect on where the algorithm converged. That is, the final estimates obtained in the combination algorithm were those that the original EM algorithm converged to as well. In practice one would try a range of values for determining when to switch algorithms. This approach can be combined with any of the EM variants as well. We coupled the ECME algorithm with the BFGS algorithm for the simulation studies, which resulted in tremendous savings in computation time. For example, using a Unix Ultra 1 workstation with 1024 MB of RAM and a 400 MHz processor, the time required to fit 500 simulated datasets was on the order of five days for the original EM algorithm, while on the order of 12 hours for the combination ECME-BFGS algorithm.

Discussion

There are a number of other factors that can influence the convergence rate of the algorithm, or even whether the algorithm converges at all. The initial starting values for the regression parameter, mass points, and masses often determine if the algorithm will converge. Initial estimates for the regression parameters can be obtained by fitting a model that excludes the random effect. One could also use the maximum likelihood estimates from one of the approaches in the previous chapter, such as the pseudo likelihood approach or full ML approach. There have been a number of suggestions for determining initial estimates of the masses and mass points. Wood and Hinde (1987) and Follmann and Lambert (1989) suggested starting values for the mass points based on the histogram of the residuals from the fixed effects model. Aitkin (1996, 1999) suggested using the weights and nodes from Gauss-Hermite quadrature as initial estimates for the masses and mass points, respectively. Butler and Louis (1992) generated initial estimates of the mass points from $\{\text{logit}(\frac{v}{K+1}),\ v = 1, \cdots, K\}$. No single method will guarantee convergence to the global maximum. It is best to try more than one set of starting values to verify that the global maximum has been reached. From our own experience, we have had success using the starting values proposed by Aitkin (1996, 1999) and Butler and Louis (1992).

When applying the NPML algorithm, convergence must be obtained within a fit when the support size K is fixed, and then between successive fits when the support size is increased. Convergence within a fit can be determined by monitoring the change in parameter estimates and the change in deviance (Aitkin 1996, 1999). If one is using the alternative MCEM-BFGS algorithm, convergence is determined within the quasi-Newton algorithm where changes in the log-likelihood and gradients of the log-likelihood are monitored. When convergence is obtained within a fit, the support size is increased and the model is refit. There are a number of ways to determine if the optimal support size has been reached. Typically an increase in the support size beyond the optimal value leads to multiplicities in mass points, or masses with zero probabilities. In conjunction with these occurrences, there is usually little to no change in the deviance between the successive fits. Thus one can determine convergence in K by comparing deviances between fits. Occasionally, however, an increase in the support size beyond the optimal value will lead to singular matrices within the Fisher Scoring algorithm as mass points take on identical values. The deviance would then be undefined, but the choice of K would be obvious.

An alternative method for finding the optimal K is by overfitting the support size (Aitkin 1996, 1999). With this approach, one successively reduces K from some large starting point until the true K is reached. We utilized this approach in the simulation studies in Section 4.5. It has been our experience, and the experience of others (Wood and Hinde 1987; Follmann and Lambert 1989; Aitkin 1996, 1999), that the optimal support size K is typically small, falling somewhere between two and six, even when the true mixing distribution is continuous. One notable exception was reported by Davies and Pickles (1987), who analyzed primary and secondary shopping behavior for 275 households in England. Davies and Pickles (1987) proposed a model based on the inverse Gaussian density and the negative exponential density, which included two nuisance parameters. They assumed that the bivariate distribution of the nuisance parameters was concentrated at a finite set of two-dimensional mass points. Using an NPML approach to maximize the log-likelihood, Davies and Pickles (1987) found that the NPML estimate required 18 mass points to fully characterize the nonparametric, bivariate mixing distribution. In Chapter 5 we discuss the extension of the existing algorithm to the bivariate case.

As noted by Wood and Hinde (1987) and from our own experience, it is possible that the maximum likelihood estimate of the mixing distribution has positive probability at $m = \pm\infty$. Indeed, in their NPML algorithm for binary response data, Wood and Hinde (1987) included mass parameters at $m = \pm\infty$ by default. To do so, they assumed that mixing distribution G was of the form

$$\omega_1\,\delta(m=-\infty) + \omega_2\,\delta(m=\infty) + \sum_{k=1}^{K} p_k\,\delta(m=m_k)$$

where $\omega_1 + \omega_2 + \sum_{k=1}^{K} p_k = 1$ and $\delta(m = m_k)$ denotes a mass point at $m = m_k$. The occurrence of mass points at plus or minus infinity depends on the distribution of the response profiles for the clusters, with clusters having all the same response contributing to the event. In a dataset of voting histories where subject's voting profiles were recorded over time, Wood and Hinde (1987) interpreted a mass point at $\pm\infty$ as those people who are certain to vote (or not to vote) for a particular party. As inclusion of such parameters in the model would remove the model from the generalized linear model framework, we did not consider such additions. Our own experience of mass points at plus or minus infinity occurred only in simulated datasets where they, in those instances, had no adverse effect on the estimates of the fixed regression parameters. Standard errors, however, were greatly influenced as they are obtained through inversion of the observed information matrix. Having an extremely large (in absolute value) mass point at the least caused the variance estimate to be negative for that mass point. When this occurred, surprisingly, the standard errors of the fixed regression parameters were still reasonable. Other times, however, the observed information matrix became uninvertable leaving no estimates of the standard errors for the parameters. If this occurred in the analysis of an actual dataset, one might wish to fit a model such as that used by Wood and Hinde (1987), which can account for mass points at plus or minus infinity.

A final observation that we have made concerns the incorporation of the mass points into the fixed effects design matrix. Recall that this was accomplished by expressing the K mass points as a set of $K-1$ dummy variables along with the fixed covariates. An alternative is to include all K dummy variables and then fit a no intercept model. We have found that this approach often leads to faster convergence rates. One reason for this is the mass point estimates are able to move about more freely, no longer being estimated as changes from the intercept/baseline mass point.

4.2.2 Inference

We now consider inference within the context of the NPML approach. Ideally, one would like to have an asymptotic theory for the estimation of the fixed effects parameters that would parallel the standard maximum likelihood theory. This would provide one with a variety of inferential techniques, such as information matrix calculations of standard errors, as well as the asymptotic justification for the use of these methods. Unfortunately such asymptotic theory is still lacking, which is a major deficiency of the NPML approach. The difficulty in deriving the asymptotic theory arises from the unknown support size for the mixing distribution. When the support size is unknown, the dimension of the parameter vector is also unknown making large sample results difficult to obtain. Despite this, the majority of the recommendations for making inference on the fixed effects parameters are based on the standard maximum likelihood theory. In this section we consider the calculation of standard error estimates for the fixed parameters and the mixing parameters in addition to likelihood inference for the NPML approach. We obtain standard errors through the calculation of the observed information matrix, with the required second derivatives given below. We then consider hypothesis testing for the fixed parameters and the mixing distribution as well as for model comparisons.

Standard errors

There have been a number of suggestions for obtaining standard errors within the context of the NPML approach (Follmann and Lambert 1989; Butler and Louis 1992; Dietz and Bohning 1995; Aitkin 1996, 1999). Follmann and Lambert (1989) and Butler and Louis (1992) obtained standard errors by calculating the observed information matrix, which we will discuss in detail below. Dietz and Bohning (1995) suggested three alternative estimators for the standard errors which they called the Profile Likelihood (PL) estimator, the Likelihood-ratio (LR) estimator, and the Multiple-Imputation (MI) estimator.

Let β denote a single parameter and $\hat{\beta}$ its maximum likelihood estimator. Also let $l_{\beta=w}$ be the supreme of the log-likelihood function given that $\beta = w$. The PL estimator is primarily an estimator of confidence intervals. Consider the values β^+ of β that fall inside the 95 percent confidence interval defined by the condition $2(l_{\beta=\hat{\beta}} - l_{\beta=\beta^+}) < \chi^2_{0.95}$. The bounds on the interval can be calculated by computing the profile likelihood for an appropriate grid of β values. Dietz and Bohning (1995) argued that if the PL confidence interval is compact and approximately symmetric with respect to $\hat{\beta}$, the standard error of $\hat{\beta}$ can be approximated by

$$\mathrm{SE}_{PL}(\hat{\beta}) = \frac{\hat{\beta}_U - \hat{\beta}_L}{2 * 1.96}, \tag{4.8}$$

where $\hat{\beta}_U$ and $\hat{\beta}_L$ are the upper and lower bound of the PL interval, respectively. Dietz and Bohning (1995) noted that the compactness and symmetry requirements are commonly met if the sample size is sufficiently large.

The second estimator, the LR estimator, is based on the property that in large samples from models for which the log-likelihood is quadratic in the parameters, the likelihood-ratio and Wald tests for the significance of an individual parameter are equivalent (Dietz and Bohning 1995). For testing a single parameter $\beta = 0$, the likelihood-ratio and Wald statistics are given by

$$\lambda_{LR} = -2(l_{\beta=0} - l_{\beta=\hat{\beta}}) \quad \text{and} \quad \lambda_W = \frac{\hat{\beta}^2}{\text{Var}(\hat{\beta})}, \tag{4.9}$$

respectively. Since the Wald test and likelihood-ratio test are equivalent asymptotically, one can equate the two equations in (4.9) to yield the standard error estimate

$$\text{SE}_{LR}(\hat{\beta}) = \left[\frac{\hat{\beta}^2}{2(l_{\beta=\hat{\beta}} - l_{\beta=0})} \right]^{1/2} . \tag{4.10}$$

This approach is also advocated by Aitkin (1996, 1999). Even if the log likelihood is skewed where equivalence of the Wald and likelihood-ratio tests may not hold, Aitkin (1996, 1999) argued that estimate (4.10) is still more appropriate than an estimate based on the inverse information matrix. He reasoned that use of (4.10) in the Wald statistic, λ_W, given in (4.9) leads to the likelihood-ratio statistic, λ_{LR}, which reflects the skewness of the log-likelihood. Using the inverse information matrix estimate of the standard error in λ_W would be misleading as it would not account for the skewed log likelihood.

The final estimator suggested by Dietz and Bohning (1995) is the Multiple-Imputation (IM) estimator. In this approach, one augments the original data with simulated membership data denoting which of the $\hat{K}$ classes the particular observation came from. The data are simulated using the parameter estimates and estimated weights α_{ik} from the final iteration of the NPML algorithm. For each of m such simulated samples, the maximum likelihood estimate and standard error of β are estimated. Denoting the standard error estimate for the qth sample as σ_q, the MI estimator for the variance of $\hat{\beta}$ is given by

$$\text{Var}_{MI}(\hat{\beta}) = \frac{2}{m} \sum_{i=1}^{m} \sigma_i^2. \tag{4.11}$$

Dietz and Bohning (1995) reported a small simulation study comparing the three approaches for calculating standard errors. They concluded that all three provided adequate estimates, though noting that the LR estimator tended to be conservative.

A major deterrent for the use of the PL, LR, and MI estimators is the additional computation required to obtain the estimates. To obtain the profile confidence interval for the PL approach, one must fit repeated models over a grid of possible β's to determine the upper and lower confidence bounds. Such calculations would require an extremely fast algorithm such as the directional derivative approach of Lesperance and Kalbfleisch (1992). The LR estimator also requires additional model fitting. For each parameter, one must fit the reduced model excluding that parameter to calculate the standard error. This approach is implemented by the statistical package GLIM4, which Aitkin (1996, 1999) has used for fitting his NPML models and may be a reason why he advocated the LR estimator (4.10) approach. The MI estimator involves an additional m fits after obtaining the maximum likelihood values for the parameters. Dietz and Bohning (1995) used $m = 1000$ for their simulation study though they do not comment on the amount of time needed to fit the m samples. Even with the current computing power, we would guess this would require an exorbitant amount of time.

For the NPML algorithm we have proposed, we obtain estimates of the standard errors by calculating the observed information matrix. As the NPML algorithm is based on the EM algorithm, one method for finding the observed information matrix is Louis' method (Louis 1982), as outlined in Chapter 3. Butler and Louis (1992) utilized this approach when they performed their simulation studies. Alternatively, one can calculate the observed information matrix directly, by evaluating at the maximum likelihood estimates the inverse of the second derivative of the log-likelihood function. We now provide the necessary derivatives for such calculations.

For notational convenience, we denote the log-likelihood function as

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \log \sum_{k=1}^{K} p_k \, f_{ik}, \tag{4.12}$$

where $f_{ik} = \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ijk} \mid \boldsymbol{\beta}^*)$. Then, letting $\boldsymbol{\Psi}' = (\boldsymbol{\beta}^{*'}, \mathbf{p}')$ be the vector of parameters, the observed information matrix is

$$F_O(\boldsymbol{\Psi}) = -\frac{d^2 l}{d\boldsymbol{\Psi} d\boldsymbol{\Psi}'} = \begin{bmatrix} F_O^{\boldsymbol{\beta}^*\boldsymbol{\beta}^*} & F_O^{\boldsymbol{\beta}^*\mathbf{p}} \\ F_O^{\mathbf{p}\boldsymbol{\beta}^*} & F_O^{\mathbf{pp}} \end{bmatrix}. \tag{4.13}$$

For the log-likelihood function (4.12),

$$\frac{dl}{dp_k} = \sum_{i=1}^{n} \frac{f_{ik} - f_{iK}}{\sum\limits_{l=1}^{K} p_l \, f_{il}} \tag{4.14}$$

and

$$\frac{dl}{d\boldsymbol{\beta}^*} = \sum_{i=1}^{n} \sum_{k=1}^{K} \alpha_{ik} \, \mathbf{s}_{ik}(\boldsymbol{\beta}^*), \tag{4.15}$$

where $\alpha_{ik} = \dfrac{p_k \, f_{ik}}{\sum_{l=1}^{K} p_l \, f_{il}}$, and $\mathbf{s}_{ik}(\boldsymbol{\beta}^*)$ denotes the contribution to the score function for the ith cluster in the kth component. Recall that the form of the score function for a particular link function is given in Chapter 2. Given (4.14) and (4.15), and denoting the contribution to the observed information matrix for the ith cluster in the kth component by $F_{O,ik}$ as defined in Chapter 2, the elements of (4.13) defined as follows. The $K-1$ by 1 vector, $F_O^{\boldsymbol{\beta}^*\mathbf{p}}$, has elements

$$F_O^{\boldsymbol{\beta}^* p_k} = -\frac{d^2 l}{d\boldsymbol{\beta}^* dp_k} = -\sum_{i=1}^{n} \left\{ \left(\frac{\alpha_{ik}}{p_k} s_{ik} - \frac{\alpha_{iK}}{p_K} s_{iK} \right) - \frac{dl}{dp_k} \frac{dl}{d\boldsymbol{\beta}} \right\},$$

$F_O^{\mathbf{pp}}$ is a $K-1$ by $K-1$ matrix with (r,s)th element

$$F_O^{p_r p_s} = -\frac{d^2 l}{dp_r dp_s} = -\sum_{i=1}^{n} \frac{dl}{dp_r} \frac{dl}{dp_s},$$

and

$$F_O^{\boldsymbol{\beta}^*\boldsymbol{\beta}^*} = -\frac{d^2 l}{d\boldsymbol{\beta}^* d\boldsymbol{\beta}^{*'}} = -\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\alpha_{ik}\, F_{O,ik} + \alpha_{ik}\,\mathbf{s}_{ik}\,\mathbf{s}_{ik}^{'} - \alpha_{ik}\left(\sum_{l=1}^{K}\alpha_{il}\,\mathbf{s}_{il}\right)\mathbf{s}_{ik}^{'},\right\}.$$

Upon convergence of the NPML algorithm, the observed information matrix is evaluated at the maximum likelihood estimates of the fixed parameters and mixing distribution and then inverted to obtain an estimated variance-covariance matrix for the parameters. We note that missing from the observed information matrix is the support size parameter K. Follmann and Lambert (1989) used a small simulation study to show that calculation of the standard errors by assuming that $\hat{K}$ was the true support size is appropriate even if $\hat{K} \neq K$. They argued that the variability in an estimated parameter $\hat{\beta}$ depends on the variability in the mixing distribution G which can be captured by $\hat{G}$ even if $\hat{K} \neq K$. In their simulation they compared Monte Carlo estimates of standard errors to those obtained from the observed information matrix where the true mixture was normal and a two point mixture. Even in the normal case where the estimated K is far from the true continuous distribution, the two estimates of the standard errors were close. In our simulation study in Section 4.5.1, we examine the performance of the observed information matrix standard errors as well, under a variety of mixture distributions.

Hypothesis testing and model comparisons

If the support size K were known, hypothesis testing for the fixed parameters and model comparisons could be carried out using the standard likelihood-ratio test as defined in (4.9). For the application of mixtures at hand, the support size is an unknown parameter and must be estimated from the data. The asymptotic distribution of the likelihood-ratio test in this context is still unknown. Consider the problem of testing a single fixed parameter equal to zero. In a fixed effects model, the distribution of the likelihood-ratio statistic would be χ^2 with one degree freedom, since the difference in parameters between the null and the alternative hypotheses is one. In the models considered here, the difference in parameters between the null and the alternative need not be one. If, for instance, an additional mass parameter was needed under the alternative hypothesis, the difference in parameters would be three: one for the fixed parameter, one for the mass point m, and one for the mass p. In this instance, one could view the additional mass p in the alternative as being zero in the null hypothesis, however the asymptotic theory for the likelihood-ratio test fails again as $p = 0$ is on the boundary of the parameter space.

Despite the lack of theoretical justification for its use, the likelihood-ratio test has been commonly used for testing fixed parameters and making model comparisons (Davies 1987; Wood and Hinde 1987; Aitkin 1996, 1999). Surprisingly, justification for its use has generally come from a single simulation study by Davies (1987). Davies (1987) suspected that the flexibility of the NPML approach in accounting for variation between sampled individuals might make it difficult to detect systematic variation due to explanatory variables. He conjectured that the power of the likelihood-ratio test would be lower for the NPML approach as compared with equivalent tests in parametric approaches. To test this claim, Davies (1987) conducted a simulation study using a negative binomial model. The negative binomial distribution is commonly used for modeling Poisson response data that exhibit overdispersion. Given that a count $y_i \mid \lambda_i$ is distributed Poisson with mean λ_i, the negative binomial model is obtained by assuming the rate parameter λ_i follows a Gamma distribution with parameters a and b. The resulting marginal distribution of y_i is negative binomial with mean $a\,b$ and variance $a(b+1)\frac{1}{b}$. Davies (1987) fixed a and b, and assumed that the count y_i was dependent on a single covariate. He then simulated from the true negative binomial model 500 datasets with the coefficient of the covariate $\beta = 0$ and 100 datasets for each of $\beta = 0.1, 0.2, \cdots, 0.7$. Davies (1987) tabulated the rejection rates for testing the null hypothesis of $\beta = 0$ from both the parametric and nonparametric fits. He concluded that there was no evidence of the likelihood-ratio test being distributed differently for mass point models. In addition, any loss of power through the use of a nonparametric approach was too small to be detected by the simulation study. Davies (1987) noted that the support sizes ranged from three to five throughout the simulation study. However, it was not clear whether the number of mass points changed under the null and alternative hypothesis for a given dataset. In Section 4.5.2 we conduct a similar simulation study to investigate the rejection rate of the likelihood-ratio test for the NPML approach compared to the parametric approach of the previous chapter.

An alternative to the likelihood-ratio test for testing a fixed parameter is the Wald test, given in (4.9). Under the usual regularity conditions (Rao 1973, p. 364), the estimates of the parameters, being maximum likelihood estimates, are asymptotically normal. Given an estimate of the asymptotic covariance matrix, as defined in the previous section, the ratio of the square of the parameter estimate and the variance of the estimator has, asymptotically, a χ^2 distribution with one degree of freedom. Use of the Wald statistic in the context of the models considered here has been limited, due to the lack of exact asymptotic standard errors. Though this is true, one could argue that use of the Wald statistic is no more incorrect than the use of the likelihood-ratio statistic. The Wald statistic is also unaffected by the potential difference in support sizes between the null and alternative hypotheses as in the likelihood-ratio test. Thus, we also examine the performance of the Wald test in the simulation study in Section 4.5.2.

As in fixed parameter hypothesis testing, the ability to make formal quantitative inferences concerning the mixing distribution G is well behind our ability to estimate it and use it informally. Though there has been considerable research in the area, many of the basic properties of the maximum likelihood estimate of G are still unknown. Certain results concerning G have been established, but only under special circumstances. For example, the asymptotic distribution of $\hat{G}$ is unknown, however Lindsay (1989) showed that the moments of G can be estimated with the usual root-n asymptotics. Of main interest for the application of mixtures considered here, is to test the heterogeneity model versus one not including a random effect. In terms of the parameters of the mixing distribution, one would test that the mixing proportions are zero. As before, this entails testing that the masses are on the boundary of the parameter space which precludes the use of the likelihood-ratio test. By simulation, Bohning et al. (1994) showed for mixtures of densities from the one-parameter exponential family that the distribution of the likelihood-ratio statistic for testing homogeneity versus a two component mixture was similar to a mixture of χ_2^2, χ_1^2, and χ_0^2. Even so, a number of authors use the change in deviance between the homogeneity and heterogeneity models as a guideline for making such decisions (Aitkin 1996, 1999).

4.3 Identifiability

When considering any mixtures of probability distributions, the problem of identifiability is of great importance. A given mixture is identifiable if it is uniquely characterized in the sense that two distinct sets of parameters defining the mixture can not yield the same distribution. In this section we consider the identifiability of the nonparametric mixture models proposed in the previous sections. Specifically, we give sufficient conditions for the identifiability of the mixture multinomial random effects model, when the mixing is over a single multinomial distribution. Such models can be considered as overdispersed multinomial models. We have already considered such a model in Section 3.6.2 where we fit a shifted threshold model to the toxicity dataset. Recall that a multinomial response vector was recorded for each litter of mice. By fitting a shifted threshold model, we accounted for possible overdispersion among the multinomial response vectors of the litters. The results shown in this section are based on extensions of results by Teicher (1963, 1967). An excellent review of identifiability and its many applications can be found in Prakasa Rao (1992).

Our goal in this section is to provide conditions under which overdispersed multinomial models are identifiable. An explicit definition of identifiability in the context of these models and conditions for their identifiability will be given in Section 4.3.2. To arrive at such conditions we first consider the identifiability of mixtures of multinomial distributions and the mixtures of products of multinomial distributions in Section 4.3.1. These results are extensions of results by Teicher (1963, 1967) where mixtures of binomial distributions were considered. As in the case of binomial mixtures, the class of all finite mixtures of multinomial distributions is not identifiable. We will show, however, that certain subsets of this family are identifiable when certain conditions are met. To determine which subsets are identifiable, we first prove Theorem 4.1, a generalization of Proposition 3 in Teicher (1963), which characterizes identifiable mixtures of general multinomial distributions. This leads directly to necessary and sufficient conditions for the identifiability of mixtures of multinomial distributions with fixed sample sizes. We then show that under certain conditions, products of multinomial distributions are also identifiable.

4.3.1 Mixtures of Multinomial and Product Multinomial Distributions

In this section we consider the identifiability of mixtures of multinomial distributions and products of multinomial distributions. We begin with an example which will illustrate the concept of identifiability.

Example 4.1

In this example we show that mixtures of two binomial distributions with sample size two are not identifiable. Let $B(2, \pi)$ denote the binomial distribution with two trials and success probability π, $0 < \pi < 1$. Let $G_{\pi_1, \pi_2, \alpha}$ be a mixing distribution with $P(\pi = \pi_1) = \alpha = 1 - P(\pi = \pi_2)$, where $\pi_1 \neq \pi_2$ and $0 < \alpha < 1$. Let X denote *a random variable with the distribution that is a mixture of $B(2,\pi)$ with respect to the mixing distribution $G_{\pi_1,\pi_2,\alpha}$. Then*

$$P_{\pi_1,\pi_2,\alpha}(X=0) = \alpha\,(1-\pi_1)^2 + (1-\alpha)(1-\pi_2)^2, \tag{4.16}$$

$$P_{\pi_1,\pi_2,\alpha}(X=1) = 2\alpha\pi_1(1-\pi_1) + 2\alpha\pi_2(1-\pi_2), \tag{4.17}$$

and

$$P_{\pi_1,\pi_2,\alpha}(X=2) = \alpha\pi_1^2 + (1-\alpha)\pi_2^2.$$

Since $\sum_{i=0}^{2} P_{\pi_1,\pi_2,\alpha}(X=i) = 1$ only equations (4.16) and (4.17) are needed to determine $P_{\pi_1,\pi_2,\alpha}(X=i)$ for $i = 0,1,2$.

The mixture of two binomials $B(2,\pi_1)$ and $B(2,\pi_2)$ with respect to $G_{\pi_1,\pi_2,\alpha}$ is said to be identifiable if

$$P_{\pi_1,\pi_2,\alpha}(X=0) = P_{\pi_1^*,\pi_2^*,\alpha^*}(X=0) \;\; \text{and} \;\; P_{\pi_1,\pi_2,\alpha}(X=1) = P_{\pi_1^*,\pi_2^*,\alpha^*}(X=1) \tag{4.18}$$

$$\text{implies that} \;\; (\pi_1,\pi_2,\alpha) = (\pi_1^*,\pi_2^*,\alpha^*). \tag{4.19}$$

Since (4.16) and (4.17) are a set of two equations with three unknowns (π_1,π_2,α), there are infinitely many solutions for a given pair of values for $P_{\pi_1,\pi_2,\alpha}(X=0)$ and $P_{\pi_1,\pi_2,\alpha}(X=1)$. For example, two such solutions that yield the same set of values are $(\pi_1,\pi_2,\alpha) = (0.5, 0.6, 0.5)$ and $(\pi_1,\pi_2,\alpha) = (0.3, 0.56, 0.038462)$. Therefore, since (4.18) does not imply (4.19), the family of mixtures of two binomials $B(2,\pi_1)$ and $B(2,\pi_2)$ is not identifiable.

Example 4.1 is a special case of a result proven by Teicher (1963). Teicher (1963) showed that for the family of binomial distributions with fixed sample size n and parameter π, a necessary and sufficient condition for the class of all finite mixtures of at most k elements of the family to be identifiable is that $n \geq 2k-1$. In Example 4.1, n and k were both two and so $n \not\geq 2k-1$. The proof of this result follows a similar argument to that used in Example 4.1. In general, condition (4.18) states

Table 4.1: Total number of multinomial vectors $\mathbf{y}$ for multinomial sample size n and number of probabilities R

	R Probabilities			
n	2	3	4	5
1	2	3	4	5
2	3	6	10	15
3	4	10	20	35
4	5	15	35	70
5	6	21	56	126
10	11	66	286	1001

that the two mixtures of binomials are the same for all possible values, x, of the binomial random variable X. Teicher (1963) expressed this condition in terms of the moments of the binomial distribution and as a function of its parameters, and showed that this system of equations has a unique solution if $n \geq 2k - 1$. We will use this same approach to prove Theorem 4.1, which extends the result of Teicher (1963) to multinomial distributions. For the multinomial case, however, the system of equations based on the moments of the multinomial distribution no longer has $n + 1$ equations. The number of equations will depend on the multinomial sample size n, but not in the simple form $n + 1$ as in the binomial case. The reason is that condition (4.18) must now hold for all possible vectors $\mathbf{y}' = (y_1, \cdots, y_q)$ of the multinomial random variable $\mathbf{Y}$ with $R = q + 1$ probabilities and sample size n, where the vector $\mathbf{y}$ is an element of $\aleph_n = \{\mathbf{y} : \mathbf{y}' = (y_1, \cdots, y_q),\ \sum_{i=1}^{q} y_i \leq n,\ y_i \in \{0, 1, 2, \cdots\},\ 1 \leq i \leq q\}$. The number of equations will then be the total number of possible vectors $\mathbf{y}$ for a given multinomial sample size n. Table 4.1 lists the number of equations for a small set of multinomial sample sizes n and number of probabilities R. For $R = 2$, the binomial case, the number of equations is again seen to be $n + 1$. In general, the number of possible equations, C_{nq}, for a multinomial sample size of n with $R = q + 1$

probabilities can be found from

$$C_{nq} = \sum_{i=0}^{n} \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} \cdots \sum_{l_{q-2}=0}^{l_{q-3}} (l_{q-2}+1), \tag{4.20}$$

where $\sum_{l_1=0}^{i} \cdots \sum_{l_{q-2}=0}^{l_{q-3}} (l_{q-2}+1)$ is defined to be 1 for $q = 1$ and $(i+1)$ for q=2.

We begin in Theorem 4.1 by characterizing the subfamilies of the class of all finite mixtures of multinomial distributions that are identifiable. Letting $\mathcal{P} = \{\boldsymbol{\pi} : \sum_{i=1}^{q} \pi_i \leq 1,\ 0 < \pi_i < 1,\ 1 \leq i \leq q\}$ and $\mathbb{N}$ be the positive integers, we denote the multinomial distribution with parameters n and $\boldsymbol{\pi}$ as $M(\mathbf{y}; n, \boldsymbol{\pi})$, where $\boldsymbol{\pi} \in \mathcal{P}$, $\mathbf{y} \in \aleph_n$, and $n \in \mathbb{N}$. In Theorem 4.1 we consider mixtures of general multinomial distributions in which both n and $\boldsymbol{\pi}$ are considered as parameters. The conclusions of Theorem 4.1 lead directly to necessary and sufficient conditions for the identifiability of mixtures of multinomial distributions with fixed n.

Theorem 4.1

Let $\mathcal{F}_1 = \{M(\mathbf{y}; n_i^{'}, \boldsymbol{\pi}_i^{'}),\ \boldsymbol{\pi}_i^{'} \in \mathcal{P},\ n_i^{'} \in \mathbb{N},\ 1 \leq i \leq k^{'}\}$ and $\mathcal{F}_2 = \{M(\mathbf{y}; n_i^{''}, \boldsymbol{\pi}_i^{''}),\ \boldsymbol{\pi}_i^{''} \in \mathcal{P},\ n_i^{''} \in \mathbb{N},\ 1 \leq i \leq k^{''}\}$ denote two finite families of multinomial distributions and let k be the number of elements in $\mathcal{F}_1 \bigcup \mathcal{F}_2$. Denote the h unique multinomial sample sizes in $\mathcal{F}_1 \bigcup \mathcal{F}_2$ as $\bar{n}_1 > \bar{n}_2 > \cdots > \bar{n}_h$ and let r_i be the number of occurrences of $\bar{n}_i$, $1 \leq i \leq h$, in $\mathcal{F}_1 \bigcup \mathcal{F}_2$.

(i) *A necessary condition for*

$$\sum_{i=1}^{k^{'}} c_i^{'} M(\mathbf{y}; n_i^{'}, \boldsymbol{\pi}_i^{'}) \underset{\forall \mathbf{y}}{\equiv} \sum_{i=1}^{k^{''}} c_i^{''} M(\mathbf{y}; n_i^{''}, \boldsymbol{\pi}_i^{''}),\ \sum_{i=1}^{k^{'}} c_i^{'} = \sum_{i=1}^{k^{''}} c_i^{''} = 1,\ 0 < c_i^{'}, c_i^{''} \tag{4.21}$$

to imply

$$k^{'} = k^{''},\quad (n_i^{'}, \boldsymbol{\pi}_i^{'}) = (n_{j_i}^{''}, \boldsymbol{\pi}_{j_i}^{''}) \tag{4.22}$$

for some permutation $(j_1, \cdots, j_k)$ *of* $(1, \cdots, k)$ *is that*

$$C_{\bar{n}_h q} \geq r_h. \tag{4.23}$$

(ii) *A sufficient condition that* (4.21) *imply* (4.22) *is that* (4.23) *and*

$$C_{\bar{n}_i q} - C_{\bar{n}_{i+1} q} \geq r_i, \quad 1 \leq i \leq h-1 \tag{4.24}$$

hold.

Remark. When (4.21) implies (4.22), the class of all finite mixtures of $k^* = k' = k''$ elements of $\mathcal{F} = \{M(\mathbf{y}; n, \boldsymbol{\pi}), \boldsymbol{\pi} \in \mathcal{P}, n \in \mathbb{N}\}$ is said to be *identifiable.*

Proof

First note that (4.21) *is equivalent to*

$$\sum_{i=1}^{k} d_i \, M(\mathbf{y}; n_i, \boldsymbol{\pi}_i) \underset{\forall \mathbf{y}}{\equiv} 0, \quad \sum_{i=1}^{k} d_i = 0, \tag{4.25}$$

where $M(\mathbf{y}, n_i, \boldsymbol{\pi}_i)$, $1 \leq i \leq k$, *are the elements of* $\mathcal{F}_1 \bigcup \mathcal{F}_2$. *Let* $s_j = \sum_{i=1}^{j} r_i$ *and* $s_0 = 0$. *Without loss of generality, we order the multinomial distributions in* (4.25) *such that* $d_1, \cdots, d_{s_1}$ *correspond to the* r_1 *distributions with sample size* $\bar{n}_1$, $d_{s_1+1}, \cdots, d_{s_2}$ *correspond to the* r_2 *distributions with sample size* $\bar{n}_2$, *and so on. The moment generating function for the multinomial distribution is*

$$M_{\mathbf{Y}}(\mathbf{t}) = \left(1 - \sum_{j=1}^{q} \pi_j + \sum_{j=1}^{q} \pi_j e^{t_j}\right)^n = \left(1 + \sum_{j=1}^{q} \pi_j w_j\right)^n$$

where $w_j = e^{t_j} - 1$. *Since the moment generating function of a random variable uniquely determines the distribution function of the random variable,* (4.25) *can be expressed as*

$$\sum_{i=1}^{k} d_i \left(1 + \sum_{j=1}^{q} \pi_{ij} w_j\right)^{n_i} \underset{\forall \mathbf{w}}{\equiv} 0,$$

which, by the multinomial theorem, leads to

$$\sum_{i=1}^{k} d_i \left\{ \sum_{\mathbf{y}_i \in \aleph_{n_i}} \frac{n_i!}{\prod_j y_{ij}!} (w_1 \pi_{i1})^{y_{i1}} \cdots (w_q \pi_{iq})^{y_{iq}} \right\} \underset{\forall \mathbf{w}}{\equiv} 0, \tag{4.26}$$

where we define $\prod_j y_{ij}! = (n_i - \sum_{l=1}^{q} y_{il})! \times \prod_{j=1}^{q} y_{ij!}$. *Careful grouping of terms in* (4.26) *leads to the following sets of equations*

$$\frac{\bar{n}_1!}{\prod_j k_j!} \sum_{i=1}^{s_1} d_i \pi_{i1}^{k_1} \cdots \pi_{iq}^{k_q} = 0, \ \mathbf{k} \in \mathcal{A}_1 \tag{4.27}$$

$$\frac{\bar{n}_1!}{\prod_j k_j!} \sum_{i=1}^{s_1} d_i \pi_{i1}^{k_1} \cdots \pi_{iq}^{k_q} + \frac{\bar{n}_2!}{\prod_j k_j!} \sum_{i=s_1+1}^{s_2} d_i \pi_{i1}^{k_1} \cdots \pi_{iq}^{k_q} = 0, \ \mathbf{k} \in \mathcal{A}_2 \tag{4.28}$$

$$\vdots$$

$$\frac{\bar{n}_1!}{\prod_j k_j!} \sum_{i=1}^{s_1} d_i \pi_{i1}^{k_1} \cdots \pi_{iq}^{k_q} + \cdots + \frac{\bar{n}_h!}{\prod_j k_j!} \sum_{i=s_{h-1}+1}^{s_h} d_i \pi_{i1}^{k_1} \cdots \pi_{iq}^{k_q} = 0, \ \mathbf{k} \in \mathcal{A}_h, \tag{4.29}$$

where $\mathcal{A}_i = \{\mathbf{y} : \bar{n}_{i+1} + 1 \leq \sum_{j=1}^{q} y_j \leq \bar{n}_i, \ y_i \in \{0, 1, 2, \cdots\}, \ 1 \leq i \leq q\}$.

We prove (**i**) *by contradiction. Assume that* (4.21) *implies* (4.22) *but* (4.23) *does not hold, that is that* $C_{\bar{n}_h q} < r_h$. *By choosing* $d_i = 0$, $1 \leq i \leq s_{h-1}$, *one can elicit a counter-example in which* (4.27)–(4.29) *and* (4.21) *hold, but* (4.22) *does not. Such a counter-example was given in Example* 4.1 *where* $(k', k'') = (n_1', n_2') = (n_1'', n_2'') = (2, 2)$ *and* $(d_1, d_2, d_3, d_4) = (.5, .5, -.038462, -1.038462)$. *Thus* (**i**) *holds.*

To see that (**ii**) *is true, note that when* (4.23) *and* (4.24) *hold, for each* $\tau = 1, \cdots, h$ *and* $\mathcal{A}_{(\cdot)}$ *defined as before*

$$\frac{\bar{n}_\tau!}{\prod_j k_j!} \sum_{i=s_{\tau-1}+1}^{s_\tau} d_i \pi_{i1}^{k_1} \cdots \pi_{iq}^{k_q} = 0, \quad \mathbf{k} \in \mathcal{A}_\tau$$

is a set of $E_\tau = C_{\bar{n}_\tau q} - C_{\bar{n}_{\tau+1} q}$ *equations with* r_τ *unknowns, where* $E_\tau \geq r_\tau$. *Thus the only solution for each set of equations is the zero solution. That is* $d_i = 0$, $1 \leq i \leq k$, *which satisfies* (4.21) *and implies* (4.22) *since* $c_i', c_i'' > 0$.

□

An immediate consequence of Theorem 4.1, paralleling Proposition 4 of Teicher (1963), is the following:

Corollary 4.1

Let $\mathcal{F} = \{M(\mathbf{y}; n, \boldsymbol{\pi}),\ \boldsymbol{\pi} \in \mathcal{P}\}$ denote a family of multinomial distributions with parameter $\boldsymbol{\pi}$ and fixed sample size n. A necessary and sufficient condition that the class of all finite mixtures of at most k elements of $\mathcal{F}$ be identifiable is that $C_{nq} \geq 2k$.

Proof

In Theorem 4.1 let $\mathcal{F}_1 = \{M(\mathbf{y}; n, \boldsymbol{\pi}_i^{'}),\ \boldsymbol{\pi}_i^{'} \in \mathcal{P},\ 1 \leq i \leq k^{'}\}$ and $\mathcal{F}_2 = \{M(\mathbf{y}; n, \boldsymbol{\pi}_i^{''}),\ \boldsymbol{\pi}_i^{''} \in \mathcal{P},\ 1 \leq i \leq k^{''}\}$. Since the only distinct multinomial sample size in $\mathcal{F}_1 \bigcup \mathcal{F}_2$ is n, $\bar{n}_1 = \bar{n}_h = n$ and $r_h = k^{'} + k^{''}$. From (**i**) *in Theorem 4.1 a necessary condition for* (4.21) *to imply* (4.22) *is that $C_{nq} \geq (k^{'} + k^{''}) = 2k$ (by* (4.22)). *But this also obtains sufficiency since condition* (4.24) *holds trivially when all multinomial sample sizes are the same.*

□

Applying Corollary 4.1 and (4.20) for $q = 1$, one obtains the result of Teicher (1963) that finite mixtures of binomial distributions are identifiable if and only if $n \geq 2k - 1$. Thus mixtures of binomials with sample size one are not identifiable. On the contrary, mixtures of multinomials with sample size one and $R > 3$ are identifiable. Using (4.20) one can see, for example, that for $q = 3, 4$ the number of identifiable mixture components for mixtures of multinomials with sample size one is two.

Teicher (1967) considered the identifiability of general mixtures of products of distributions. Let $\mathcal{F}_1^* = \{F(\mathbf{y}; \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathcal{R}^m\}$ represent a family of distributions $F(\mathbf{y}; \boldsymbol{\alpha})$ and let $\mathcal{F}_n^* = \{F^*(\mathbf{y}; \boldsymbol{\alpha}) : F^*(\mathbf{y}; \boldsymbol{\alpha}) = \prod_{i=1}^n F(\mathbf{y}_i; \boldsymbol{\alpha}_i), F(\mathbf{y}_i; \boldsymbol{\alpha}_i) \in \mathcal{F}_1^*,\ \ 1 \leq i \leq n\}$ so that if $\mathbf{Y}_1, \cdots, \mathbf{Y}_n$ are independent random variables each of whose distributions is in $\mathcal{F}_1^*$, their joint distribution is an element of $\mathcal{F}_n^*$. Teicher (1967) showed the following:

Theorem 4.2 (Theorem 2, Teicher [1967])

If the class of all finite mixtures of $\mathcal{F}_1^*$ is identifiable, then for every $n > 1$ the class of finite mixtures of $\mathcal{F}_n^*$ is likewise identifiable.

Using Theorem 4.2 and Corollary 4.1, we conclude this section by giving conditions under which mixtures of products of multinomial distributions are identifiable.

Theorem 4.3

Let $\mathcal{F}_i = \{M(\mathbf{y}; n_i, \boldsymbol{\pi}_i),\ \boldsymbol{\pi}_i \in \mathcal{P}\},\ \ 1 \le i \le N,$ be N families of multinomial distributions with parameter $\boldsymbol{\pi}_i$ and fixed sample size n_i and let $\mathcal{F}_N = \{F(\mathbf{y}; \mathbf{n}, \boldsymbol{\pi}) : F(\mathbf{y}; \mathbf{n}, \boldsymbol{\pi}) = \prod_{i=1}^{N} M(\mathbf{y}; n_i, \boldsymbol{\pi}_i),\ \ M(\mathbf{y}; n_i, \boldsymbol{\pi}_i) \in \mathcal{F}_i,\ \ 1 \le i \le N\}$. Then the class of all finite mixtures of at most k elements of $\mathcal{F}_N$ is identifiable if $\min_{1\le i\le N} \{C_{n_iq}\} \ge 2k$.

Proof

By Corollary 4.1, each class of all finite mixtures of at most k elements of $\mathcal{F}_i$ is identifiable if and only if $C_{n_iq} \ge 2k,\ \ 1 \le i \le N$. Let $C_{(1)} = \min_{1\le i\le N}\{C_{n_iq}\}$. Clearly Corollary 4.1 holds simultaneously for all N families when $C_{(1)} \ge 2k$. Therefore, applying Theorem 4.2, the class of all finite mixtures of at most k elements of $\mathcal{F}_N$ is identifiable if $C_{(1)} \ge 2k$.

□

As in Theorem 4.1, finite mixtures of products of distribution functions that are identifiable satisfy the following uniqueness of representation property:

$$\sum_{i=1}^{k'} c_i' \prod_{j=1}^{N} M(\mathbf{y}; n_j, \boldsymbol{\pi}_{ij}') \underset{\forall \mathbf{y}_N^*}{\equiv} \sum_{i=1}^{k''} c_i'' \prod_{j=1}^{N} M(\mathbf{y}; n_j, \boldsymbol{\pi}_{ij}''), \tag{4.30}$$

$$\sum_{i=1}^{k'} c_i' = \sum_{i=1}^{k''} c_i'' = 1,\ 0 < c_i', c_i''$$

implies

$$k' = k'',\ \ (\boldsymbol{\pi}_{i1}', \cdots, \boldsymbol{\pi}_{iN}') = (\boldsymbol{\pi}_{l_i1}'', \cdots, \boldsymbol{\pi}_{l_iN}'')$$

for some permutation $(l_1, \cdots, l_k)$ of $(1, \cdots, k)$. The equality in (4.30) holds for all $\mathbf{y}_N^{*'} = (\mathbf{y}_1', \cdots, \mathbf{y}_N')$ where $\mathbf{y}_j \in \aleph_{n_j}$ is applied to the jth multinomial in (4.30), $1 \leq j \leq N$.

4.3.2 Mixtures of Multinomial Regression Models

Using results of the previous section, we now consider the identifiability of finite mixtures of overdispersed multinomial logit regression models. There has been relatively little research on the identifiability of mixtures of logistic regression models for binary responses. Wood and Hinde (1987) mention that such models are identifiable if at least one regression variable is continuous and are identifiable only in rare circumstances when the covariates are all discrete. They do not, however, provide conditions under which particular models would be identifiable. For finite mixtures of logistic regression models with random intercepts, Follmann and Lambert (1991) provide sufficient conditions that ensure that the regression parameters and the mixing distribution of the intercept are identifiable. For mixtures of binomials with sample sizes greater than one, they apply the results of Teicher (1963) to bound the number of components in the mixing distribution. For mixtures of Bernoulli distributions, they estimate the mixing distribution from distributions with similar covariate values. In this case, the bound on the number of components of the mixing distribution depends on the number of covariate vectors that agree on all coordinates except for one. For both sets of conditions, the logistic regression models that they consider allow for mixtures of one binomial or one Bernoulli distribution. More recently Butler and Louis (1997) considered a latent linear model for binary data having the structure of (1.1) with any class of mixing distribution. They give sufficient conditions for identifiability of the fixed effects and mixing distribution as well as for convergence of their maximum likelihood estimators. Though their results are general and can be applied to a wide range of models, the conditions for assuring identifiability are quite esoteric and do not provide information as to the number of mass points that are identifiable.

We begin by defining identifiability within the context of overdispersed multinomial logit regression models. As in Section 4.2, we consider models of the form $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + u_i$ where the random effect u_i is assumed to have a discrete distribution G with finite masses $p_1 > 0, \cdots, p_K > 0$, $\sum_{k=1}^{K} p_k = 1$, and mass points $m_1 < \cdots < m_K$. For overdispersed multinomial logit models, the number of observations per cluster is one ($T_i = 1$, $i = 1, \cdots, n$), thus we drop the j subscript for the remainder of this section. We assume that conditional on the random effect u_i, the multinomial observation $\mathbf{y}_i$, with corresponding probability vector $\boldsymbol{\pi}_i$, has distribution

$$f(\bar{\mathbf{y}}_i \mid Z_i, n_i, \boldsymbol{\beta}, u_i),$$

where n_i denotes the multinomial sample size. The parameter $\boldsymbol{\pi}_i$ in the multinomial distribution can take any of the forms given in Chapter 2. The mixed, unconditional probability function for $\mathbf{y}_i$ is given by

$$f(\bar{\mathbf{y}}_i \mid Z, n_i, \boldsymbol{\beta}, G) = \sum_{k=1}^{K} p_k \; f(\bar{\mathbf{y}}_i \mid Z_i, n_i, \boldsymbol{\beta}, m_k). \tag{4.31}$$

As one of the m_k is aliased with one of the thresholds, we assume that the column corresponding to the first threshold in Z_i has been removed to allow direct estimation of all K mass points. Let $\mathcal{G}_{K'}$ be the set of all discrete distributions on $(-\infty, \infty)$ with a support size of at most K' points. For given (Z_i, n_i), $i = 1, \cdots, n$, we define the set of parameters $\{(G, \boldsymbol{\beta}) : G \in \mathcal{G}_{K'}, \boldsymbol{\beta} \in \mathbb{R}^p\}$ to be identifiable if

$$\begin{gathered} f(\bar{\mathbf{y}}_i \mid Z_i, n_i, \boldsymbol{\beta}, G) \equiv f(\bar{\mathbf{y}}_i \mid Z_i, n_i, \boldsymbol{\beta}^*, G^*) \\ \text{for all } \; \mathbf{y}_i \text{ and } \; 1 \leq i \leq n \end{gathered} \tag{4.32}$$

implies that

$$(G, \boldsymbol{\beta}) = (G^*, \boldsymbol{\beta}) \;\text{ for } G, G^* \in \mathcal{G}_{K'} \;\text{ and }\; \boldsymbol{\beta}, \boldsymbol{\beta}^* \in \mathbb{R}^p,$$

where G and G^* have support size $K \le K'$ and $K^* \le K'$, respectively.

Using Theorem 4.3 we now give a sufficient condition for the identifiability of model (4.31) by bounding the support size K. This generalizes Theorem 1 of Follmann and Lambert (1991) which provides conditions for the identifiability of binary logistic regression models where the mixing is over a single binomial distribution. Loosely stated, we use the cluster with the largest multinomial sample size to identify the mixing distribution, while the remaining clusters are used to identify the regression parameters.

Theorem 4.4

Consider the set of finite mixed multinomial regression models defined in equation (4.31) where $\mathbf{z}_{ir} \in \mathbb{R}^p$ for $r = 1, \cdots, q$, $i = 1, \cdots, n$, and $Z_i = [\mathbf{z}_{ir}]$. Let the index I be such that $\mathcal{C}_I = \max\{\mathcal{C}_{n_1,q}, \cdots, \mathcal{C}_{n_n,q}\}$, where $C_{n_i,q}$ is defined as in (4.20). If each set of vectors $\{\mathbf{z}_{1r} - \mathbf{z}_{Ir}, \cdots, \mathbf{z}_{I-1,r} - \mathbf{z}_{Ir}, \mathbf{z}_{I+1,r} - \mathbf{z}_{Ir}, \cdots, \mathbf{z}_{nr} - \mathbf{z}_{Ir}\}$, $r = 1, \cdots, q$, spans $\mathbb{R}^p$ then $\{(G, \boldsymbol{\beta}) : G \in \mathcal{G}_{K'}, \boldsymbol{\beta} \in \mathbb{R}^p\}$ is identifiable for $K' \le \frac{1}{2}\mathcal{C}_I$.

Proof

Consider distributions G and G^* in $\mathcal{G}_{K'}$ that satisfy equation (4.32). Then the mixed multinomial distribution with parameters $(n_I, \boldsymbol{\pi}_I(\boldsymbol{\beta}))$ and mixing distribution G, and the mixed multinomial distribution with parameters $(n_I, \boldsymbol{\pi}_I(\boldsymbol{\beta}^*))$ and mixing distribution G^* are identical. Since finite mixtures of multinomials without covariates are identifiable if $K \le K'$ (Corollary 4.1), $G = G^*$ and $\pi_{Ir}(m_k + \mathbf{z}_{Ir}'\boldsymbol{\beta}) = \pi_{Ir}(m_k^* + \mathbf{z}_{Ir}'\boldsymbol{\beta}^*)$ for $r = 1, \cdots, q$ and $k = 1, \cdots, K'$. Because $\pi(\cdot)$ is monotone, $m_k^* = m_k + \mathbf{z}_{Ir}'\boldsymbol{\beta} - \mathbf{z}_{Ir}'\boldsymbol{\beta}^*$, for $r = 1, \cdots, q$ and $k = 1, \cdots, K'$.

Since equality of distributions implies equality of means,

$$n_i \sum_{k=1}^{K'} p_k \, \pi_{ir}(m_k + \mathbf{z}_{ir}'\boldsymbol{\beta}) = n_i \sum_{k=1}^{K'} p_k \, \pi_{ir}(m_k + \mathbf{z}_{Ir}'(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \mathbf{z}_{ir}'\boldsymbol{\beta}^*),$$

$r = 1, \cdots, q$. *Again using the monotonicity of* $\pi(\cdot)$,

$$\begin{aligned} m_k + \mathbf{z}_{Ir}'\boldsymbol{\beta} &= m_k + \mathbf{z}_{Ir}'(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \mathbf{z}_{ir}'\boldsymbol{\beta}^* \\ 0 &= \mathbf{z}_{ir}'(\boldsymbol{\beta} - \boldsymbol{\beta}^*) - \mathbf{z}_{Ir}'(\boldsymbol{\beta} - \boldsymbol{\beta}^*) \\ 0 &= (\mathbf{z}_{ir} - \mathbf{z}_{Ir})'(\boldsymbol{\beta} - \boldsymbol{\beta}^*), \end{aligned} \tag{4.33}$$

for $r = 1, \cdots, q$, $i = 1, \cdots, n$. *Finally, because of the assumption that* $\{\mathbf{z}_{1r} - \mathbf{z}_{Ir}, \cdots, \mathbf{z}_{I-1,r} - \mathbf{z}_{Ir}, \mathbf{z}_{I+1,r} - \mathbf{z}_{Ir}, \cdots, \mathbf{z}_{nr} - \mathbf{z}_{Ir}\}$, $r = 1, \cdots, q$, *spans* $\mathbb{R}^p$, *condition* (4.33) *holds only if* $(\boldsymbol{\beta} - \boldsymbol{\beta}^*) = 0$, *or* $\boldsymbol{\beta} = \boldsymbol{\beta}^*$. *Hence, it follows that* $m_k = m_k^*$, $k = 1, \cdots, K'$. □

In Theorem 4.4, we assumed that a common parameter vector $\boldsymbol{\beta}$ held for all logits. The theorem still holds, however, if separate parameter vectors $\boldsymbol{\beta}_r$ are allowed for each logit. As an application of Theorem 4.4, we consider the toxicity dataset found in Table 3.3. Theorem 4.4 bounds the identifiable support size by the largest $\frac{1}{2}C_{n_i,q}$ over all clusters. For the toxicity dataset, the largest multinomial sample is 16. Using the definition of $C_{n_i,q}$ in (4.20), the maximum identifiable number of components for a multinomial sample size of 16 with three response probabilities is 76. Thus, one need not worry about identifiability when fitting the shifted threshold model to this dataset. As noted before, Theorem 4.4 can be used when a single multinomial observation is observed for each cluster. It is also important to establish similar conditions when more than one multinomial observation is collected for each cluster. One might be able to accomplish this using Theorem 4.3, which establishes identifiability of products of multinomial distributions. More research is still needed in this area.

4.4 Application

We now apply the NPML algorithm of Section 4.2.1 to the wine dataset considered in the previous chapter. As before, we simply illustrate the NPML estimation method as opposed to providing a thorough analysis. We also assume that the models are identifiable. For the bitterness of wine dataset, we apply the NPML algorithm using both the cumulative logit and adjacent-category logit links to fit the shifted threshold model. From the analysis of the toxicity data and satisfaction data in the previous chapter, it was clear that the shifted threshold model was inadequate and that the vary threshold model was more appropriate. Though we motivated the NPML algorithm for the shifted threshold model, it can be modified to fit the varying threshold model as well. We will consider such modifications in Chapter 5.

4.4.1 Cumulative Logit Link

Consider again the data in Table 3.2 consisting of ratings of wines with respect to bitterness. Recall that each of nine judges rated eight wines on a five-point scale ranging from least to most bitter. Factors in the experiment included temperature (TE) of the wine (cold/warm), whether there was contact (CO) with the skin when the grapes were crushed (yes/no), and bottle (BO) number (first/second). As each wine judge has a particular sensitivity to the bitterness of wine, one would expect their responses to be correlated. We again model the heterogeneity among judges by allowing the thresholds to be shifted for each judge. However, we now assume that the random effect follows a discrete mixing distribution. We first consider a cumulative logit model with linear predictor

$$\eta_{ijr} = \alpha_r + \beta_{TE} x_{ij1} + \beta_{CO} x_{ij2} + \beta_{BO} x_{ij3} + u_i, \qquad (4.34)$$

$$r = 1, \cdots, R-1, \; j = 1, \cdots, T, \; i = 1, \cdots, n,$$

where $R = 5$, $T = 8$, and $n = 9$. In (4.34) β_{TE}, β_{CO}, and β_{BO} are the parameter coefficients for the temperature, contact, and bottle factors, respectively. As in the analysis in Section 3.5.1, the factors were coded 1 and -1 to correspond to the original analysis by Tutz and Hennevogl (1996).

Using the algorithm discussed in Section 4.2.1, we calculated maximum likelihood estimates for model (4.34) by successively increasing the support size K. Table 4.2 contains the NPML estimates for support sizes of two, three, and four, as well as the maximum likelihood estimates obtained using 5-point adaptive Gauss-Hermite (AGH(5)) quadrature where the random effect was assumed to be normal. For the NPML approach, we fit a no-intercept version of model (4.34) to obtain direct estimates of the mass points. An estimate of the suppressed threshold, α_1, was obtained from the mean of the mixing distribution

$$\hat{\mu}_{\hat{K}} = E[u_i] = \sum_{k=1}^{\hat{K}} \hat{m}_k \hat{p}_k. \tag{4.35}$$

To obtain the correct estimates of the remaining thresholds, the estimates of α_2, α_3, and α_4 from the original fit of model (4.34) must by increased by $\hat{\mu}_{\hat{K}}$. The estimates of the standard deviation of the mixing distribution given in Table 4.2 were calculated using the standard variance formula for discrete random variables

$$\hat{\sigma}^2_{\hat{K}} = V[u_i] = \sum_{k=1}^{\hat{K}} \hat{m}^2_k \hat{p}_k - \hat{\mu}^2_{\hat{K}}. \tag{4.36}$$

It is clear from Table 4.2 that the NPML estimate of the mixing distribution has a support size of three. The log-likelihood values for support sizes of two, three, and four are -82.583, -80.237, and -80.237, respectively, indicating that increasing the mass point from three to four is unnecessary. In addition, mass point two from the model with three mass points is repeated when K is increased to four. Thus the estimated mixing distribution has mass points -1.522, -4.023, and -6.007, with respective masses .113, .676, and .212. By centering the mixing distribution about

Table 4.2: Nonparametric maximum likelihood analysis of model (4.34) for the wine bitterness dataset using the cumulative logit link. Estimates were obtained for support sizes of two, three, and four. The final column contains the estimates using 5-point adaptive Gauss-Hermite (AGH(5)) quadrature under the assumption of a normal random effect.

	Nonparametric ML			ML
	Estimated Support Size ($\hat{K}$)			
Parameter	2	3	4	AGH(5)
α_1	-3.820	-4.161	-4.161	-4.082
α_2	-0.898	-0.952	-0.952	-0.930
α_3	1.663	1.840	1.840	1.797
α_4	3.498	3.763	3.763	3.657
β_{TE}	1.468	1.562	1.562	1.536
	(0.292)	(0.300)	(0.301)	(0.298)
β_{CO}	0.862	0.938	0.938	0.916
	(0.251)	(0.262)	(0.256)	(0.256)
β_{BO}	0.115	0.124	0.124	0.122
	(0.230)	(0.235)	(0.236)	(0.232)
σ	0.934	1.232	1.232	1.145
Mass Point m_1	-3.189	-1.522	-1.522	
Mass p_1	0.687	0.113	0.113	
Mass Point m_2	-5.203	-4.023	-4.023	
Mass p_2	0.313	0.676	0.379	
Mass Point m_3		-6.007	-4.023	
Mass p_3		0.212	0.297	
Mass Point m_4			-6.007	
Mass p_4			0.212	
Log-likelihood	-82.583	-80.237	-80.237	-81.394

its mean, the centered mass points are found to be 2.639, 0.138, and -1.846. Thus the mixing distribution is somewhat symmetric about zero.

We now compare the results from the three point NPML fit to the adaptive quadrature results. We can see that the estimates for all parameters are very similar between the two approaches. The NPML parameter estimates and standard errors are consistently larger, but only marginally so. The standard deviation estimate based on the three point discrete distribution was 1.232, compared with 1.145 under the assumption of normality. Statistical conclusions regarding the significance of the effect parameters would be the same for both approaches. We note that the maximized log-likelihood value for the NPML approach was slightly larger (-80.237) than the adaptive quadrature approach (-81.394). This usually occurs as the NPML algorithm has a greater number of parameters due to the masses and mass points.

4.4.2 Adjacent-Category Logit Link

We likewise used the NPML algorithm to fit model (4.34) with the adjacent-category logit link. Recall that the adjacent-category logit model provides odds ratio estimates for adjacent pairs of responses. Thus, interpretations are based on a subset of the response scale as opposed to the entire response scale as in the cumulative logit model. Parameter estimates for support sizes of two, three, and four are found in Table 4.3 along with the adaptive quadrature results reported in Section 3.6.1.

We again see that only a three-point mixing distribution is needed to obtain the NPML estimates. Note that with a support size of four we obtain a redundancy in the first mass point (1.381). Parameter estimates and standard errors for the three-point NPML results and the adaptive quadrature results are in close agreement. Slightly larger values are obtained for the NPML approach, as was seen in Table 4.2. Both approaches would yield the same statistical conclusions.

Table 4.3: Nonparametric maximum likelihood analysis of model (4.34) for the wine bitterness dataset using the adjacent-category logit link. Estimates were obtained for support sizes of two, three, and four. The final column contains the estimates using 8-point adaptive Gauss-Hermite (AGH(8)) quadrature under the assumption of a normal random effect.

	Nonparametric ML			ML
	Estimated Support Size ($\hat{K}$)			
Parameter	2	3	4	AGH(8)
α_1	0.000	-0.016	-0.016	-0.009
α_2	-0.947	-1.117	-1.117	-1.043
α_3	-1.718	-2.008	-2.008	-1.880
α_4	-2.037	-2.415	-2.415	-2.250
β_{TE}	-1.067	-1.219	-1.219	-1.149
	(0.246)	(0.275)	(0.275)	(0.268)
β_{CO}	-0.614	-0.700	-0.700	-0.659
	(0.192)	(0.210)	(0.210)	(0.203)
β_{BO}	-0.052	-0.059	-0.059	-0.056
	(0.161)	(0.172)	(0.172)	(0.167)
σ	0.662	0.971	0.971	0.839
Mass Point m_1	1.097	1.381	1.381	
Mass p_1	0.267	0.228	0.001	
Mass Point m_2	-0.399	-0.145	1.381	
Mass p_2	0.733	0.661	0.227	
Mass Point m_3		-2.110	-0.145	
Mass p_3		0.111	0.661	
Mass Point m_4			-2.110	
Mass p_4			0.111	
Log-likelihood	-81.683	-79.495	-79.495	-80.853

4.5 Simulation Studies

Since the underlying distribution of the random effect is typically unknown, the nonparametric approach provides a potentially robust alternative to the models discussed in Chapter 3. In this section we present two simulation studies in which we compare the NPML approach to that of the ML approach considered in the previous chapter. In the first simulation study we compare the bias in parameter estimates between the two approaches for a single covariate, random intercept cumulative logit model using a variety of different random effects distributions. In the second study we examine the behavior of the likelihood-ratio and Wald test statistics for testing a fixed effect parameter by comparing the rejections rates between the NPML and ML approaches.

4.5.1 Simulation Study I

A number of authors have used simulation studies to explore the behavior of the NPML and ML approaches. Follmann and Lambert (1989) conducted a small simulation study to test their conjecture that the knowledge of K was unimportant for estimating standard errors. They considered two overdispersed logistic regression models each with a single binary covariate in which they assumed the mixing distribution was either a normal distribution or a two point discrete distribution. From their simulations they concluded that calculating standard errors by treating the estimated support size K as the true value did not adversely affect the standard error estimates. Neuhaus et al. (1992) investigated the effects of misspecification of the mixing distribution when fitting random intercept logistic regression models using the ML approach of the previous chapter. Using the normal, gamma, and t-distribution as mixing distributions, they measured the bias in the parameter estimates and standard errors when one assumed the mixing distribution was normal. They concluded that the ML approach was quite robust to misspecification of the mixing distribution. However, inferences concerning the mixing distribution was much less robust to model misspecification.

Butler and Louis (1992) compared parametric and nonparametric approaches for random effects models in both linear and logistic models. They utilized simulation to study the moderate and large-sample performance of the NPML method to that of the Gaussian method within the context of random intercept logistic regression models. For all simulations they considered only the standard normal distribution for the true mixing distribution. They concluded that the NPML method performed efficiently when the true mixing distribution was Gaussian. Recently Aitkin and Alfo (1998) proposed a set of conditional models for modeling binary longitudinal responses which mixes features of random effects models and transition models and allows for a general association structure between different observations on the same individual. They presented both a parametric estimation method, based on Gauss-Hermite quadrature, and an NPML estimation method for fitting the proposed models. To assess the practical performance of their proposed models, Aitkin and Alfo (1998) conducted a large simulation study using a first-order Markov chain model with a single covariate and random effect. In the simulation experiment they varied the number of subjects analyzed, the number of observations per subject, the true parameter values for the two covariates, and they considered six possible distributions for the single random effect. Though their main objective was to compare their proposed set of models, Aitkin and Alfo (1998) noted that misspecification of the number of components of the mixing distribution generally caused bias in only the autoregressive term of the model and not in the covariate term.

We conducted a set of simulations to compare the performance of the nonparametric approach to that of the normal random effects approach for datasets with clustered ordinal outcomes and varying mixture distributions. For the simulations we generated clustered ordinal response data according to a cumulative logit model with linear predictor

$$\eta_{ijr} = \alpha_r + \beta\, x_{ij} + u_i, \quad r = 1, \cdots, R-1, \tag{4.37}$$

where the number of responses R=3, $i = 1, \cdots, n$, and $j = 1, \cdots, T$. For all simulations, the $\{x_{ij}\}$ in model (4.37) were generated from the standard normal distribution and the fixed parameter vector $(\alpha_1, \alpha_2, \beta) = (-1, 1, 0.5)$. We considered two sets of simulations, a larger set in which the number of clusters, n, was fixed at 100, and smaller set where $n = 1000$. In the larger set we varied the number of observations per cluster T = (4 or 7) and the true distribution of the random effect u_i. We considered five cases for the mixture distribution: $N(0, \sigma^2 = 0.5)$, $Exp(1)$, $U(-0.5; 0.5)$, discrete with mass points (-0.5,0.5) and masses (0.5,0.5), and no random effect. For each case except the last, the random effect was standardized to have mean zero and variance 0.5. In the smaller set of simulations, only a subset of the cluster sizes and mixture distributions were considered due to the computational burden of using 1000 clusters. Here we fixed the cluster size at seven and consider only the normal and exponential distributions. The two simulation sets can be summarized as follows. The numbers inside the table indicate the total number of simulations run for each algorithm.

	SET I		SET II
CLUSTER SIZE:	4	7	7
NORMAL	500	500	100
EXPONENTIAL	500	500	100
UNIFORM	500	500	
DISCRETE	500	500	
NO RANDOM EFFECT	500	500	

To assess the performance of the ML and NPML approaches, Monte Carlo estimates of the bias in parameter estimates for each approach were calculated where the bias for $\hat{\theta}$ is defined as: Bias $= E(\hat{\theta}) - \theta$. For $n = 100$ a pilot study was used to determine that 500 simulations were needed to yield Monte Carlo standard error estimates of less than 0.01. Thus using 500 simulations at each of the ten combinations of cluster size and mixture distribution, estimates of the bias of the fixed parameters $(\alpha_1, \alpha_2, \beta)$ and the variance of the mixing distribution σ^2 were calculated by fitting model (4.37) using the NPML approach and the ML approach of the previous chapter. For the simulation set using 1000 clusters, only 100 simulations were run for the normal and exponential distributions. As in Follmann and Lambert (1989), average standard error estimates for $\hat{\beta}$ were calculated from both the observed information matrix at each model fit and from the Monte Carlo estimate across all simulations.

For the ML approach, we used 15-point adaptive Gauss-Hermite quadrature to directly maximize the log-likelihood. Starting values for the fixed effect parameters were obtained by fitting the fixed effects logistic regression model while the starting value of 0.75 was used for σ^2. For each model fit, the standard errors of the parameters were calculated by evaluating the analytical observed information matrix. For the NPML approach, the ECME-BFGS algorithm was used to obtain estimates of the fixed effects and mixing distribution. The estimates of the mass points and masses were then used to estimate σ^2. At each simulation, the support size K had to be estimated as well. We accomplished this by overfitting the number of mass points (Aitkin 1996). For each combination we began with K fixed at seven and successively reduced K until convergence was obtained. In order to determine convergence we needed to define some stopping rules. Upon convergence at a given K, mass points that were less than 0.0001 apart were combined and K was reduced accordingly. Also, if any masses were less than 0.0001, they, and their corresponding mass point, were removed and K adjusted accordingly. If at K no further points could be removed, an additional run was made at $K-1$ and log-likelihoods compared to ensure that convergence had been met. Upon convergence standard errors were calculated from the analytical observed information matrix. Initial estimates of the fixed parameters were obtained in the same manner as in the ML approach. The initial estimates of the mass points were obtained from $\{\text{logit}(\frac{v}{K+1}),\ v = 1, \cdots, K\}$ while the masses were set at $1/K$.

For some simulated datasets, the NPML algorithm attempted to converge to mass points that were at plus or minus infinity. In such instances the mass point estimate continues to grow to plus or minus infinity with little increase in the overall log-likelihood value. Thus the algorithm would signify convergence with the offending mass point having an extremely large absolute value. Since reducing the cluster size yielded a smaller log-likelihood fit, we did not replace these simulations. However, such extreme estimates usually produced problems when calculating the observed information matrix. Thus we report two sets of tables with and without the problem simulations.

The results of the first set of simulations are found in Tables 4.4 – 4.22, grouped by random effects distribution and cluster size. As found by Neuhaus et al. (1992), there was little estimated bias in the estimation of the regression coefficient β for the parametric maximum likelihood approach. This was true even when the mixing distribution was extremely skewed (i.e. the exponential distribution, Tables 4.12 and 4.14). The largest estimated biases in β, 0.013 and 0.012, occurred for cluster sizes of four for both the uniform random effect (Table 4.10) and with no random effect (Table 4.21), respectively, and these could be explained by Monte Carlo error. This corresponds to a percent bias (Bias/True Value) of approximately 2.5%, which is similar to the percent biases reported by Neuhaus et al. (1992). In general, the simulations with cluster size four tended to exhibit larger bias in β then the corresponding simulations with cluster size seven. Across all simulations there was strong agreement between the simulation and model based estimates of the standard error for β. This suggests that valid variance estimates of the estimated covariate effects can be obtained even under misspecification.

In contrast to the estimation of β, the estimates for the threshold parameters in the parametric approach were influenced by the skewness of the mixing distribution. The largest estimated biases occurred for the exponential distribution, regardless of the cluster size (Tables 4.12 – 4.15), and were on the order of 2%. The simulations for the remaining distributions exhibited percent biases of 0.5% to 1%. As expected, the parametric approach provided very accurate estimates of the variance component σ^2 with a normal random effect (Tables 4.4 and 4.6). There was considerable bias in σ^2 for the remaining distributions, however. In general, larger estimated biases were found for the smaller cluster size. For example, absolute percent biases of about 12% were found for the exponential and no random effect cases with cluster sizes of four (Tables 4.14 and 4.21). As noted before, the estimated standard errors from the observed information matrix for β still agreed well with the Monte Carlo estimates even when σ^2 was not well estimated.

We now turn our attention to the simulation results for the nonparametric approach. When averaged across all simulations, the estimated bias in the regression parameter β in the nonparametric approach was very similar to that of the parametric approach. It is also evident that the estimation of β was not adversely affected by mass points at plus or minus infinity. Consider Tables 4.14 and 4.15 which contain the results for the exponential distribution with cluster size of four. A large number of the simulated datasets (115) resulted in a mass point at plus or minus infinity. However the estimated biases in β with and without these simulations were only 0.006 and 0.005, respectively, approximately the level of the Monte Carlo error. The nonparametric approach exhibited the largest estimated biases with the uniform distribution and cluster size of four (percent bias of 3.6%), and with no random effect and a cluster size of four (percent bias of 2.8%). Thus, the parametric and nonparametric approaches behaved similarly in regards to the estimation of β.

Examining the standard errors for β, we see that there is close agreement between the Monte Carlo estimates and the model based estimates for the nonparametric approach as well. In addition, the standard errors from the nonparametric approach are very similar to those obtained from the parametric approach. We see that even when the simulations with mass points at plus or minus infinity are included, the model based standard errors still perform well. When broken down by the estimated support size $\hat{K}$, the standard errors tend to increase as the support size increases. This was also seen by Follmann and Lambert (1989). Comparing the results obtained for the normal random effect (Tables 4.4 – 4.7) and the 2-point discrete random effect (Tables 4.16 – 4.19), we see that being far from the true support size of the random effects distribution does not adversely effect the standard error estimates. This further supports the claim by Follmann and Lambert (1989) that knowledge of K is unimportant for estimating the standard errors.

Estimation of the remaining parameters, namely the thresholds parameters and the variance component, is considerably less accurate for the nonparametric approach. The reason for this is that these parameters are functions of the estimated mass points. Thus, when the mass points tend to plus or minus infinity, the estimates for these parameters are greatly affected. Even when the offending simulations are removed, the estimates for the variance component are generally not very good. For most of the distributions, these estimates had percent biases on the order of 5% to 10%. However, they were nearly 20% for the 2-point discrete distribution (Table 4.19) and with no random effect (Table 4.22), each having cluster sizes of four. Thus, as has been cautioned by others, one should not place too much faith in the parameter estimates that are functions of the mixing distribution parameters.

Table 4.4: Estimated bias of parameter estimates using NPML and ML estimation with a normal random effect, 100 clusters, and cluster size of seven. NPML results are listed by estimated support size and averaged across all support sizes. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	2	3	4	5	All	Likelihood
α_1	0.006	-0.001	-0.016	-0.149	-0.003	0.005
α_2	-0.003	0.009	0.031	-0.063	0.008	0.009
β	-0.007	0.006	-0.022	-0.013	0.000	-0.002
(SE($\hat{\beta}$)$_O$)	(0.076)	(0.077)	(0.076)	(0.077)	(0.076)	(0.077)
(SE($\hat{\beta}$)$_{MC}$)	(0.070)	(0.080)	(0.074)	(0.062)	(0.078)	(0.077)
σ^2	-0.084	0.213	1.3798	2.415	0.309	-0.009
Runs	111	326	56	7	500	500

Examining the occurrences of mass points at plus or minus infinity, we see that it is more likely to occur with the cluster size of four. This was to be expected since these occurrences are associated with clusters having all the same responses. This obviously will occur more often when the cluster sizes are small. We also note that the largest the support size K was estimated to be, across all of the simulations, was five. Even for the continuous random effects distributions, the support size remained between two and five. For the 2-point discrete distribution, the majority of the simulations had estimated support sizes of two, however, there were some that had estimates as high as four. Likewise, estimates of three and four were obtained when no random effect was used.

The results of the second simulation set using 1,000 clusters of size seven are found in Table 4.23, for the normal random effect, and Table 4.24 for the exponential random effect. Similar patterns are seen in these results, but with generally smaller estimated biases. It is interesting to note in Table 4.24 that the bias estimate for the variance component under the parametric approach did not improve when compared with the simulations using 100 clusters (Table 4.12). In contrast, the nonparametric estimated bias improved from 0.076 to 0.025. We also see that increasing the number

Table 4.5: Estimated bias of parameter estimates using NPML and ML estimation with a normal random effect, 100 clusters, and cluster size of seven, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. SE$(\hat{\beta})_O$ and SE$(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	2	3	4	5	All	Likelihood
α_1	0.006	-0.004	-0.009	-0.049	-0.002	0.005
α_2	-0.003	0.014	0.035	-0.062	0.013	0.009
β	-0.007	0.005	-0.017	-0.022	0.000	-0.002
(SE$(\hat{\beta})_O$)	(0.076)	(0.077)	(0.077)	(0.076)	(0.077)	(0.077)
(SE$(\hat{\beta})_{MC}$)	(0.070)	(0.080)	(0.071)	(0.074)	(0.077)	(0.077)
σ^2	-0.084	0.033	0.164	0.133	0.077	-0.009
Runs	111	321	50	5	487	500

Table 4.6: Estimated bias of parameter estimates using NPML and ML estimation with a normal random effect, 100 clusters, and cluster size of four. NPML results are listed by estimated support size and averaged across all support sizes. SE$(\hat{\beta})_O$ and SE$(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML						
	$\hat{K}$						Maximum
	1	2	3	4	5	All	Likelihood
α_1	0.099	-0.004	-0.021	-0.089	0.121	-0.014	0.000
α_2	-0.027	0.016	0.006	-0.034	0.097	0.009	0.010
β	0.035	0.000	0.001	-0.014	-0.018	0.000	-0.004
(SE$(\hat{\beta})_O$)	(0.097)	(0.104)	(0.103)	(0.092)	(0.107)	(0.103)	(0.104)
(SE$(\hat{\beta})_{MC}$)	(0.066)	(0.105)	(0.093)	(0.130)	(–)	(0.100)	(0.099)
σ^2	–	0.916	5.956	20.301	0.970	3.915	0.007
Runs	6	245	230	18	1	500	500

Table 4.7: Estimated bias of parameter estimates using NPML and ML estimation with a normal random effect, 100 clusters, and cluster size of four, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. $\mathrm{SE}(\hat{\beta})_O$ and $\mathrm{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML						
	$\hat{K}$						Maximum
	1	2	3	4	5	All	Likelihood
α_1	0.099	-0.013	0.000	-0.06	0.121	-0.006	0.000
α_2	-0.027	0.008	0.021	0.089	0.097	0.013	0.010
β	0.035	-0.001	0.005	-0.034	-0.018	0.002	-0.004
$(\mathrm{SE}(\hat{\beta})_O)$	(0.097)	(0.104)	(0.106)	(0.110)	(0.107)	(0.104)	(0.104)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(0.066)	(0.106)	(0.093)	(0.052)	(–)	(0.100)	(0.099)
σ^2	–	-0.032	0.106	0.365	0.107	0.052	0.007
Runs	6	243	146	3	1	399	500

Table 4.8: Estimated bias of parameter estimates using NPML and ML estimation with a uniform random effect, 100 clusters, and cluster size of seven. NPML results are listed by estimated support size and averaged across all support sizes. $\mathrm{SE}(\hat{\beta})_O$ and $\mathrm{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	2	3	4	5	All	Likelihood
α_1	0.003	0.070	0.167	0.438	0.056	0.013
α_2	0.009	0.075	0.165	0.530	0.061	0.009
β	0.003	0.007	0.000	0.169	0.006	0.003
$(\mathrm{SE}(\hat{\beta})_O)$	(0.079)	(0.079)	(0.076)	(0.083)	(0.079)	(0.079)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(0.081)	(0.086)	(0.071)	(–)	(0.084)	(0.083)
σ^2	-0.030	1.344	4.245	10.202	1.133	0.012
Runs	172	285	42	1	500	500

Table 4.9: Estimated bias of parameter estimates using NPML and ML estimation with a uniform random effect, 100 clusters, and cluster size of seven, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. $\text{SE}(\hat{\beta})_O$ and $\text{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML				
	$\hat{K}$				Maximum
	2	3	4	All	Likelihood
α_1	0.003	0.012	0.006	0.010	0.013
α_2	0.009	0.022	-0.010	0.016	0.009
β	0.003	0.007	0.001	0.005	0.003
$(\text{SE}(\hat{\beta})_O)$	(0.079)	(0.080)	(0.080)	(0.079)	(0.079)
$(\text{SE}(\hat{\beta})_{MC})$	(0.081)	(0.088)	(0.079)	(0.085)	(0.083)
σ^2	-0.030	0.059	0.080	0.031	0.012
Runs	172	252	25	449	500

Table 4.10: Estimated bias of parameter estimates using NPML and ML estimation with a uniform random effect, 100 clusters, and cluster size of four. NPML results are listed by estimated support size and averaged across all support sizes. $\text{SE}(\hat{\beta})_O$ and $\text{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	1	2	3	4	All	Likelihood
α_1	0.029	0.314	0.228	1.339	0.305	0.013
α_2	-0.230	0.301	0.256	1.369	0.309	0.006
β	0.035	0.013	0.025	0.008	0.018	0.013
$(\text{SE}(\hat{\beta})_O)$	(0.099)	(0.107)	(0.108)	(0.111)	(0.107)	(0.108)
$(\text{SE}(\hat{\beta})_{MC})$	(0.091)	(0.104)	(0.119)	(0.097)	(0.111)	(0.108)
σ^2	–	15.893	6.389	215.338	17.325	0.021
Runs	2	269	215	14	500	500

Table 4.11: Estimated bias of parameter estimates using NPML and ML estimation with a uniform random effect, 100 clusters, and cluster size of four, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. $\text{SE}(\hat{\beta})_O$ and $\text{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	1	2	3	4	All	Likelihood
α_1	0.029	0.019	0.001	0.126	0.011	0.013
α_2	-0.230	0.009	0.055	-0.069	0.023	0.006
β	0.035	0.014	0.035	0.013	0.021	0.013
$(\text{SE}(\hat{\beta})_O)$	(0.099)	(0.107)	(0.110)	(0.113)	(0.108)	(0.108)
$(\text{SE}(\hat{\beta})_{MC})$	(0.091)	(0.104)	(0.111)	(0.146)	(0.107)	(0.108)
σ^2	–	0.107	0.263	0.480	0.076	0.021
Runs	2	261	126	5	394	500

Table 4.12: Estimated bias of parameter estimates using NPML and ML estimation with an exponential random effect, 100 clusters, and cluster size of seven. NPML results are listed by estimated support size and averaged across all support sizes. $\text{SE}(\hat{\beta})_O$ and $\text{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML						
	$\hat{K}$						Maximum
	1	2	3	4	5	All	Likelihood
α_1	-0.013	-0.022	0.034	0.184	0.243	0.037	-0.019
α_2	-0.070	-0.027	0.049	0.215	0.254	0.048	-0.011
β	0.048	0.002	0.007	0.009	0.023	0.006	0.003
$(\text{SE}(\hat{\beta})_O)$	(0.073)	(0.075)	(0.076)	(0.077)	(0.077)	(0.076)	(0.077)
$(\text{SE}(\hat{\beta})_{MC})$	(0.002)	(0.073)	(0.079)	(0.068)	(–)	(0.076)	(0.076)
σ^2	–	-0.128	1.159	10.056	0.699	1.869	-0.054
Runs	2	136	301	60	1	500	500

Table 4.13: Estimated bias of parameter estimates using NPML and ML estimation with an exponential random effect, 100 clusters, and cluster size of seven, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML						
	$\hat{K}$						Maximum
	1	2	3	4	5	All	Likelihood
α_1	-0.013	-0.022	-0.003	-0.003	0.243	-0.008	-0.019
α_2	-0.070	-0.027	0.014	0.050	0.254	0.005	-0.011
β	0.048	0.002	0.006	0.016	0.023	0.006	0.003
(SE($\hat{\beta}$)$_O$)	(0.073)	(0.075)	(0.076)	(0.078)	(0.077)	(0.076)	(0.077)
(SE($\hat{\beta}$)$_{MC}$)	(0.002)	(0.073)	(0.080)	(0.072)	(–)	(0.077)	(0.076)
σ^2	–	-0.128	0.055	0.224	0.699	0.076	-0.054
Runs	2	136	276	38	1	453	500

Table 4.14: Estimated bias of parameter estimates using NPML and ML estimation with an exponential random effect, 100 clusters, and cluster size of four. NPML results are listed by estimated support size and averaged across all support sizes. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	1	2	3	4	All	Likelihood
α_1	-0.038	0.010	0.140	0.195	0.071	-0.023
α_2	-0.120	0.001	0.188	0.313	0.090	-0.017
β	-0.039	-0.003	0.017	0.033	0.006	0.000
(SE($\hat{\beta}$)$_O$)	(0.095)	(0.101)	(0.100)	(0.095)	(0.101)	(0.104)
(SE($\hat{\beta}$)$_{MC}$)	(0.060)	(0.104)	(0.105)	(0.142)	(0.106)	(0.105)
σ^2	–	0.491	7.011	42.637	4.688	-0.060
Runs	7	263	213	17	500	500

Table 4.15: Estimated bias of parameter estimates using NPML and ML estimation with an exponential random effect, 100 clusters, and cluster size of four, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. $\mathrm{SE}(\hat{\beta})_O$ and $\mathrm{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	1	2	3	4	All	Likelihood
α_1	-0.038	-0.010	-0.034	-0.031	-0.018	-0.023
α_2	-0.120	-0.015	0.034	0.144	0.002	-0.017
β	-0.039	-0.002	0.020	0.049	0.005	0.000
$(\mathrm{SE}(\hat{\beta})_O)$	(0.095)	(0.102)	(0.106)	(0.110)	(0.104)	(0.104)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(0.060)	(0.102)	(0.111)	(0.173)	(0.107)	(0.105)
σ^2	0.095	-0.054	0.192	0.269	0.020	-0.060
Runs	7	253	116	9	385	500

Table 4.16: Estimated bias of parameter estimates using NPML and ML estimation with a two-point discrete random effect, 100 clusters, and cluster size of seven. NPML results are listed by estimated support size and averaged across all support sizes. $\mathrm{SE}(\hat{\beta})_O$ and $\mathrm{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML				
	$\hat{K}$				Maximum
	2	3	4	All	Likelihood
α_1	0.003	-0.002	0.033	0.002	0.010
α_2	0.011	0.011	0.063	0.014	0.005
β	-0.001	0.003	0.028	0.002	0.000
$(\mathrm{SE}(\hat{\beta})_O)$	(0.076)	(0.077)	(0.075)	(0.076)	(0.077)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(0.074)	(0.074)	(0.077)	(0.074)	(0.075)
σ^2	-0.005	1.060	1.211	0.497	0.016
Runs	268	207	25	500	500

Table 4.17: Estimated bias of parameter estimates using NPML and ML estimation with a two-point discrete random effect, 100 clusters, and cluster size of seven, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. SE$(\hat{\beta})_O$ and SE$(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML				
	$\hat{K}$				Maximum
	2	3	4	All	Likelihood
α_1	0.003	-0.001	0.003	0.001	0.010
α_2	0.011	0.014	0.033	0.013	0.005
β	-0.001	0.001	0.020	0.001	0.000
(SE$(\hat{\beta})_O$)	(0.076)	(0.077)	(0.078)	(0.077)	(0.077)
(SE$(\hat{\beta})_{MC}$)	(0.074)	(0.074)	(0.075)	(0.074)	(0.075)
σ^2	-0.005	0.075	0.142	0.034	0.016
Runs	268	200	23	491	500

Table 4.18: Estimated bias of parameter estimates using NPML and ML estimation with a two-point discrete random effect, 100 clusters, and cluster size of four. NPML results are listed by estimated support size and averaged across all support sizes. SE$(\hat{\beta})_O$ and SE$(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	1	2	3	4	All	Likelihood
α_1	0.010	0.005	-0.031	-0.088	-0.011	0.003
α_2	-0.016	0.011	0.021	-0.053	0.013	0.005
β	0.168	0.012	0.016	-0.019	0.013	0.008
(SE$(\hat{\beta})_O$)	(0.100)	(0.104)	(0.104)	(0.107)	(0.104)	(0.105)
(SE$(\hat{\beta})_{MC}$)	(0.217)	(0.103)	(0.109)	(0.127)	(0.107)	(0.104)
σ^2	–	0.022	3.999	11.305	1.783	0.041
Runs	2	302	182	14	500	500

Table 4.19: Estimated bias of parameter estimates using NPML and ML estimation with a two-point discrete random effect, 100 clusters, and cluster size of four, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	1	2	3	4	All	Likelihood
α_1	0.010	0.005	-0.045	-0.098	-0.010	0.003
α_2	-0.016	0.011	0.020	-0.081	0.013	0.005
β	0.168	0.012	0.014	0.035	0.013	0.008
(SE($\hat{\beta}$)$_O$)	(0.100)	(0.104)	(0.107)	(0.108)	(0.105)	(0.105)
(SE($\hat{\beta}$)$_{MC}$)	(0.217)	(0.103)	(0.106)	(0.197)	(0.106)	(0.104)
σ^2	–	0.022	0.274	0.162	0.094	0.041
Runs	2	302	126	5	435	500

Table 4.20: Estimated bias of parameter estimates using NPML and ML estimation with no random effect, 100 clusters, and cluster size of seven. NPML results are listed by estimated support size and averaged across all support sizes. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	1	2	3	4	All	Likelihood
α_1	-0.002	-0.013	-0.048	-0.029	-0.010	-0.007
α_2	0.008	0.018	0.047	-0.025	0.015	0.012
β	-0.002	0.003	0.018	0.069	0.001	0.000
(SE($\hat{\beta}$)$_O$)	(0.072)	(0.073)	(0.075)	(0.074)	(0.073)	(0.073)
(SE($\hat{\beta}$)$_{MC}$)	(0.076)	(0.073)	(0.076)	(–)	(0.075)	(0.075)
σ^2	–	0.080	0.145	0.079	0.046	0.029
Runs	228	248	23	1	500	500

Table 4.21: Estimated bias of parameter estimates using NPML and ML estimation with no random effect, 100 clusters, and cluster size of four. NPML results are listed by estimated support size and averaged across all support sizes. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML				
	$\hat{K}$				Maximum
	1	2	3	All	Likelihood
α_1	-0.010	-0.034	-0.071	-0.024	-0.024
α_2	-0.005	0.047	0.106	0.024	0.014
β	0.002	0.023	0.062	0.014	0.012
(SE($\hat{\beta}$)$_O$)	(0.096)	(0.099)	(0.105)	(0.098)	(0.099)
(SE($\hat{\beta}$)$_{MC}$)	(0.106)	(0.105)	(0.100)	(0.106)	(0.106)
σ^2	–	1.258	1.090	0.665	0.056
Runs	234	253	13	500	500

Table 4.22: Estimated bias of parameter estimates using NPML and ML estimation with no random effect, 100 clusters, and cluster size of four, excluding simulations where a mass point was located at plus or minus infinity. NPML results are listed by estimated support size and averaged across all support sizes. SE($\hat{\beta}$)$_O$ and SE($\hat{\beta}$)$_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML				
	$\hat{K}$				Maximum
	1	2	3	All	Likelihood
α_1	-0.010	-0.046	-0.121	-0.030	-0.024
α_2	-0.005	0.040	0.090	0.019	0.014
β	0.002	0.025	0.057	0.014	0.012
(SE($\hat{\beta}$)$_O$)	(0.096)	(0.100)	(0.105)	(0.098)	(0.099)
(SE($\hat{\beta}$)$_{MC}$)	(0.106)	(0.104)	(0.105)	(0.106)	(0.106)
σ^2	–	0.176	0.418	0.098	0.056
Runs	234	224	11	469	500

of clusters from 100 to 1,000 did not change the range of the estimated support size. Between the two simulations, only one dataset resulted in a support size greater than four. This adds further support to the conjecture that the estimated support size for single random effect models will generally be small.

We can draw a number of conclusions based on the results of these simulations. First, both the parametric and nonparametric approaches exhibited similar biases when estimating the regression parameter β. In addition, the standard errors obtained for $\hat{\beta}$ from the observed information matrix were in close agreement with the Monte Carlo estimates for both approaches. They also tended to have similar behavior under the various random effects distributions and cluster sizes, both having largest estimated biases for clusters sizes of four with the uniform distribution and no random effect. For estimation of the remaining parameters, the parametric approach was much more reliable. The nonparametric approach did not provide very accurate estimates of the thresholds or the variance component. This is mainly due to mass points being estimated at plus or minus infinity. In general, the parametric approach had small estimated bias in the thresholds and variance component, except when the random effects distribution was extremely skewed. Thus, if one is interested in estimation of the thresholds or variance component, the parametric approach will generally provide more accurate estimates. For the estimation of β, however, both approaches will generally yield similar estimates.

4.5.2 Simulation Study II

As noted in Section 4.2.2, the asymptotic theory needed for making inferences in the NPML approach is still unknown. As a result of this, a number of authors have relied on the standard techniques of maximum likelihood inference for the nonparametric models, assuming that these approaches would be approximately correct (Davies 1987; Davies and Pickles 1987; Aitkin 1996, 1999). Davies (1987) provided

Table 4.23: Estimated bias of parameter estimates using NPML and ML estimation with a normal random effect, 1,000 clusters, and cluster size of seven. NPML results are listed by estimated support size and averaged across all support sizes. $\text{SE}(\hat{\beta})_O$ and $\text{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML					
	$\hat{K}$					Maximum
	2	3	4	5	All	Likelihood
α_1	-0.013	0.003	0.011	0.016	0.005	0.003
α_2	0.002	0.007	0.001	0.013	0.005	0.004
β	0.011	0.000	0.003	0.032	0.001	0.000
$(\text{SE}(\hat{\beta})_O)$	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
$(\text{SE}(\hat{\beta})_{MC})$	(0.011)	(0.025)	(0.025)	(-)	(0.025)	(0.023)
σ^2	-0.091	-0.010	-0.001	0.035	-0.009	0.000
Runs	2	66	31	1	100	100

Table 4.24: Estimated bias of parameter estimates using NPML and ML estimation with an exponential random effect, 1,000 clusters, and cluster size of seven. NPML results are listed by estimated support size and averaged across all support sizes. $\text{SE}(\hat{\beta})_O$ and $\text{SE}(\hat{\beta})_{MC}$ denote the standard error of $\hat{\beta}$ computed from the observed information matrix and from the Monte Carlo estimates across all simulations.

	Nonparametric ML				
	$\hat{K}$				Maximum
	2	3	4	All	Likelihood
α_1	-0.015	-0.002	0.021	0.002	-0.018
α_2	-0.022	0.001	0.029	0.006	-0.009
β	-0.008	-0.008	0.005	-0.007	-0.003
$(\text{SE}(\hat{\beta})_O)$	(0.025)	(0.025)	(0.018)	(0.025)	(0.025)
$(\text{SE}(\hat{\beta})_{MC})$	(0.021)	(0.023)	(0.025)	(0.022)	(0.025)
σ^2	-0.112	-0.021	0.393	0.025	-0.053
Runs	6	73	21	100	100

some evidence to support this approach in a simulation study comparing the rejection rates of the likelihood-ratio test between the NPML approach and a parametric approach. He found that the likelihood-ratio test using the NPML approach had similar type I error rate and power when compared with the likelihood-ratio test using a negative binomial model. His results have been the main justification for the use of the likelihood-ratio statistic in the NPML models (Aitkin 1996, 1999).

In this section we report the results of a simulation study that further investigates the use of standard inferential procedures in the NPML models. In contrast to Davies (1987), we examine the performance of both the likelihood-ratio test and the Wald test for the NPML approach as compared with the parametric approach of Chapter 3. Specifically, we simulated data from a cumulative logit model with linear predictor

$$\eta_{ijr} = \alpha_r + \beta x_{ij} + u_i, \quad r = 1, \cdots, R-1, \tag{4.38}$$

where the number of response, R, was three, $\alpha_1 = -1$, $\alpha_2 = .5$, and x_{ij}, $j = 1, \cdots, 7$, $i = 1, \cdots, 100$, was drawn from a standard normal distribution. Two sets of simulations were run in which we varied the true distribution of the random effect, u_i. For the first simulation set, we assumed that $u_i \sim N(0, \sigma^2 = .8)$, whereas in the second we assumed that $u_i \sim Exp(1)$. As in the first simulation study, we scaled the exponential distribution to have the same mean and variance as the normal distribution. For each simulation set, we sampled 500 datasets at each of $\beta = (0, .2, .4, .6)$ and fit both the NPML model and the cumulative logistic-normal model using adaptive Gauss-Hermite quadrature to approximate the normal integral. Simulation sizes of 500 provided Monte Carlo error estimates for the parameters of less than 0.01. To calculate the likelihood-ratio test statistic for each run, it was necessary to fit models with and without β to obtain the log-likelihood values under the null and alternative hypotheses. The Wald statistic was obtained at each run by calculating the ratio of the square of $\hat{\beta}$ and the variance of $\hat{\beta}$, obtained from the observed information matrix.

For each simulated dataset, the test of the null hypothesis, $H_o : \beta = 0$, was carried out for the likelihood-ratio test and the Wald test. The calculation of starting values and the number of quadrature points that were used in the first simulation study were also applied here.

Table 4.25 contains the estimated type I error rates for the Wald and likelihood-ratio tests for both estimating approaches, under normal and exponential random effects assumptions. It is clear from the results that the NPML and ML approaches are in close agreement with respect to their estimated type I error rates. For the normal random effect, both tests for both approaches exhibit estimated type I error rates close to the nominal level with no discernable patterns. For the exponential random effect, both tests tended to have larger estimated type I error rates than the nominal level. This held irregardless of the estimation approach. The Monte Carlo error for the estimated Wald and Likelihood-ratio tests was approximately 0.1. Thus the overestimation of the type I error rates for the exponential distribution may be spurious.

Table 4.26 contains the estimated power for the Wald and likelihood-ratio tests when testing $\beta = .2$, .4, and .6. Again we see very similar power between the parametric and nonparametric approaches for both tests. We also see a marginal increase in the power when we move from the normal random effect to the exponential random effect, which could be explained by Monte Carlo error. The close agreement found in Tables 4.25 and 4.26 between the NPML and ML approaches suggests that use of the standard asymptotic inferential tools is reasonable for the NPML approach. We can also conclude that the Wald test can be used in lieu of the likelihood-ratio test as they seem to behave similarly for the NPML approach.

As discussed before, the asymptotic theory regarding the likelihood-ratio test is known to break down when testing involves parameters on the boundary of the parameter space. This can occur with the testing of a fixed parameter using the NPML

Table 4.25: Estimated type I error rates for the Wald and likelihood-ratio (LRT) tests for testing $H_o : \beta = 0$ in a cumulative logit random intercept model. Models were fitted using both the NPML and ML approaches with both a normal and an exponential random effect distribution.

		Normal		Exponential	
LEVEL	TEST	NPML	ML	NPML	ML
0.10	LRT	0.095	0.090	0.121	0.116
	WALD	0.083	0.088	0.119	0.116
0.05	LRT	0.048	0.052	0.063	0.060
	WALD	0.054	0.050	0.061	0.060
0.01	LRT	0.006	0.006	0.016	0.012
	WALD	0.006	0.006	0.016	0.012

Table 4.26: Estimated power for the Wald and likelihood-ratio (LRT) tests for testing $H_o : \beta = .2, .4, .6$ in a cumulative logit random intercept model. Models were fitted using both the NPML and ML approaches with both a normal and an exponential random effect distribution.

		Normal				Exponential			
	TEST:	LRT		WALD		LRT		WALD	
β	LEVEL	NPML	ML	NPML	ML	NPML	ML	NPML	ML
0.2	0.001	0.232	0.226	0.219	0.218	0.253	0.254	0.247	0.244
	0.0001	0.080	0.080	0.075	0.072	0.084	0.078	0.079	0.078
	0.00001	0.024	0.030	0.021	0.022	0.032	0.034	0.023	0.032
0.4	0.001	0.956	0.960	0.954	0.956	0.968	0.964	0.959	0.964
	0.0001	0.862	0.866	0.857	0.866	0.900	0.886	0.886	0.868
	0.00001	0.776	0.774	0.757	0.754	0.751	0.762	0.735	0.726
0.6	0.001	1.000	1.000	0.996	1.000	0.973	1.000	1.000	1.000
	0.0001	1.000	1.000	0.996	1.000	0.955	1.000	1.000	1.000
	0.00001	1.000	1.000	0.996	1.000	0.940	0.998	0.998	0.998

Table 4.27: Estimated type I error rates for the Wald and likelihood-ratio (LRT) tests for testing $H_o : \beta = 0$ in a cumulative logit random intercept model. These results include only the simulations that resulted in a different support size between the null and alternative model for the likelihood-ratio test. The numbers in parentheses denote the number of simulations for each distribution.

		Normal (34)		Exponential (35)	
LEVEL	TEST	NPML	ML	NPML	ML
0.10	LRT	0.294	0.265	0.200	0.229
	WALD	0.226	0.235	0.171	0.229
0.05	LRT	0.088	0.118	0.086	0.114
	WALD	0.161	0.088	0.086	0.114
0.01	LRT	0.000	0.000	0.029	0.029
	WALD	0.000	0.000	0.029	0.029

algorithm if the support size changes under the null and alternative hypotheses. One does not have control over when this will occur, thus it is difficult to examine this by simulation. However, we examined the data obtained from the various simulations and subsetted the results into those tests that did have a differing support size between hypothesis. We do not attempt to make any strong conclusions from this data as the sample sizes are, in general, not that large. The results are given in Tables 4.27 and 4.28. Indeed in Table 4.27 the sample sizes for the normal and exponential distribution cases were only 34 and 35, thus it is difficult to make any valid conclusions. In general it seems that the likelihood-ratio test for the NPML model is performing similarly to the Wald test, and similar results seem to be obtained between approaches as well. In Table 4.28 we also see similar rejection rates both between tests and between approaches. We note that rejection rates are again seen to be higher for the exponential distribution when compared with the normal distribution. Tentatively we conclude that the likelihood-ratio test can provide approximate inferences even when the support sizes differ between the null and alternative hypotheses.

The nonparametric maximum likelihood approach for modeling clustered nominal and ordinal data is a viable alternative to the standard parametric approach discussed

Table 4.28: Estimated power for the Wald and likelihood-ratio (LRT) tests for testing $H_o : \beta = .2, .4, .6$ in a cumulative logit random intercept model. These results include only the simulations that resulted in a different support size between the null and alternative model for the likelihood-ratio test.

		Normal				Exponential			
	TEST:	LRT		WALD		LRT		WALD	
β	LEVEL	NPML	ML	NPML	ML	NPML	ML	NPML	ML
		N=57				N=69			
0.2	0.001	0.246	0.263	0.246	0.263	0.319	0.377	0.333	0.362
	0.0001	0.088	0.088	0.105	0.088	0.159	0.159	0.152	0.159
	0.00001	0.035	0.035	0.035	0.035	0.029	0.044	0.030	0.044
		N=113				N=100			
0.4	0.001	0.956	0.956	0.956	0.956	0.990	0.990	0.970	0.990
	0.0001	0.894	0.894	0.894	0.894	0.940	0.930	0.920	0.910
	0.00001	0.850	0.841	0.814	0.814	0.820	0.840	0.800	0.790
		N=150				N=252			
0.6	0.001	1.000	1.000	1.000	1.000	0.968	1.000	1.000	1.000
	0.0001	1.000	1.000	1.000	1.000	0.948	1.000	1.000	1.000
	0.00001	1.000	1.000	1.000	1.000	0.941	0.996	1.000	0.996

in Chapter 3. The EM algorithm proposed in Section 4.2 is relatively simple, and, coupled with one of the EM accelerators, can be quite fast. Though asymptotic maximum likelihood theory does not exist for the NPML approach, the simulations in this section suggest that standard tests can provide at least approximate inference. The decision of which approach to use will depend on a number of factors. If one is interested in estimation of the regression coefficients, the first simulation study suggests that both approaches will yield similar estimates and standard errors. For estimation of the mixing distribution and threshold parameters, however, the NPML approach can be quite poor if a mass point is located at plus or minus infinity. It is evident from the first simulation study that the parametric approach is quite robust to departures from normality. Thus one might consider using this approach all the time. Indeed, we have found few occurrences in which the nonparametric approach has provided substantially different results from the parametric approach. Therefore, we view the NPML approach as an additional tool for random effects modeling. It can be used as a check for the normality assumption by seeing if changes occur in the parameter estimates when the random effects distribution is estimated nonparametrically. Additional research in the areas of asymptotic theory for inferential procedures and methods for testing the number of mass points is still needed, however. In addition, software for implementing the NPML approach is also needed to allow for practical application of such models.

CHAPTER 5
METHODS FOR ANALYZING ORDINAL MULTI-CENTER CLINICAL TRIAL DATA.

5.1 Introduction

In this chapter we consider data of the form obtained from multi-center clinical trials. Clinical trials typically involve a comparison of a standard treatment versus a new treatment. Often the outcome is some binary response (success/failure). Subjects are randomly assigned to one of the two treatment groups, and the success rates of the two treatments are compared to determine which treatment is more efficacious. An important feature in many clinical trials is that the trials are carried out at multiple sites or centers. Since randomization is within center, patients from the same center often have similar treatment outcomes. This may be due, for example, to unmeasured variables such as the ability of the staff or the quality of the equipment at each center. Thus, besides just a treatment effect, there may be a center effect influencing the efficacy of the treatments. Data of this type can also arise in the area of meta-analysis where one combines results from a number of smaller studies. Such analyses are used to summarize conclusions and to better understand sources of between-study variability.

The analysis of multi-center clinical trial data has received considerable attention in the literature. Indeed, journals such as *Applied Clinical Trials* and *Controlled Clinical Trials* are dedicated to this area alone. Much attention has focused on appropriate ways of assessing the treatment effect while accounting for the possible heterogeneity in the centers and possible treatment-by-center interaction. Recent emphasis has been on the use of frequentist techniques, such as random effects models (Agresti and Hartzel 1999), and Bayesian techniques (Jones et al. 1998) for modeling this type of data. The majority of this work, however, has focused on clinical trials with binary responses.

Table 5.1 contains part of a dataset from a multi-center clinical trial conducted at Merck pharmaceutical company. The table contains the results for eight centers from a double-blind, parallel-group clinical study. The purpose of the study was to compare a new drug to a placebo for treatment of asthma. Patients were randomly assigned to a treatment and, at the end of the study, their change in asthma condition was evaluated. The response scale for this study was ordinal with possible responses of much better, better, unchanged or worse. To correctly estimate the new drug effect one must consider the effects of the individual centers and the possibility of nonconstant treatment effects among the centers. Indeed, Fleiss (1986) wrote "The most challenging questions in the analysis of the data from a multi-center trial are how to carry out the analysis when there is a treatment-by-center interaction, and, prior to that, how to ascertain whether such interaction exists." In this chapter we consider the use of the multinomial random effects models from Chapters 3 and 4 for the analysis of data such as that given in Table 5.1.

Before considering the random effects models, we first briefly discuss, in Section 5.2, fixed effects approaches for analyzing ordinal multi-center data. In Section 5.3 we consider ordinal random effects models that allow for random center effects as well as random treatment-by-center interactions. In most multi-center clinical trials the number of centers is only small to moderate in size. In Section 5.4 we report the results of a simulation study aimed at assessing the performance of the random interaction model when the number of centers is not large. To statistically determine if the random interaction term is needed, we propose a score test in Section 5.5, based on adaptive Gauss-Hermite quadrature, for testing that the variance component for a correlated random effect is zero. We use simulations to assess the performance of the proposed score test.

Table 5.1: Asthma multi-center clinical trial data from Merck pharmaceutical company comparing a drug to a placebo.

		Response		
		Much		Unchanged
Center	Treatment	Better	Better	/Worse
1	Drug	13	7	6
	Placebo	1	1	10
2	Drug	2	5	10
	Placebo	2	2	1
3	Drug	11	23	7
	Placebo	2	8	2
4	Drug	7	11	8
	Placebo	0	3	2
5	Drug	15	3	5
	Placebo	1	1	5
6	Drug	13	5	5
	Placebo	4	0	1
7	Drug	7	4	13
	Placebo	1	1	11
8	Drug	15	9	2
	Placebo	3	2	2

5.2 Fixed Effects Approach

Before discussing the fixed effects approaches for modeling ordinal multi-center data, we first consider the choice of a fixed or random effects approach. Indeed, Senn (1998) labeled this issue as one of the controversies in analyzing multi-center trials. In certain situations, the decision of a fixed or random effects approach is clear. For example, if the centers have been randomly sampled and one has an interest in estimating the center effects, a random effects approach would be appropriate. Though centers are typically not chosen at random, Grizzle (1987) argued that "...the assumption of random clinic effect will result in tests and confidence intervals that better capture the variability inherent in the system more realistically than when clinic effects are considered fixed." If, on the other hand, there are very few centers or the centers chosen are the only ones of interest, then a fixed effects analysis would be appropriate. Unlike the random effects analysis, however, the ability to estimate the center effects in the fixed effects approach will depend on the sparseness of the dataset.

Senn (1998) listed a number of pros and cons for using fixed or random effects. In support of the fixed approach he argued that the "center" is not a well defined experimental unit and that the random effects model gives it substantive significance that it does not really have. He also argued that it is difficult to precisely define what question the random effects models is answering. On the other hand, he argued that the random effects approach allows one to borrow information from all other centers when making inference on a particular center. It also provides wider and hence more realistic standard errors and confidence limits than the fixed effects analysis. One can make strong arguments for either approach as they each have their place. As our main focus in this dissertation is the application of random effects models, we only briefly consider the fixed effects approach here.

5.2.1 Maximum Likelihood

For ordinal multi-center data, such as that given in Table 5.1, the primary goal is to evaluate the efficacy of the new drug. One can accomplish this through the use of a multinomial regression model for ordinal data which includes an association parameter for the treatment effect. We only consider multinomial models based on the logit link; however, one could use alternative links, such as the probit link, as well. As before, let $\mathbf{y}_{ij}$ denote the multinomial response for the jth treatment in the ith center with multinomial sample size n_{ij}, $j = 1, 2$, $i = 1, \cdots, n$. A simple model for estimating the association between the ordinal response and the treatment covariate is

$$\eta_{ijr} = \alpha_r + \gamma_i + \beta\, x_j, \quad r = 1, \cdots, q, \tag{5.1}$$

where α_r denotes the rth threshold parameter, γ_i denotes the ith center effect, and x_j is coded one for the new drug and zero for the placebo. In order to estimate the n center effects, model (5.1) requires a constraint such as $\alpha_1 = 0$. Model (5.1) assumes a common association β holds across all centers. That is, it assumes that a treatment-by-center interaction does not exists. If one were to fit model (5.1) using a cumulative logit link, $\exp(\beta)$ would denote the common cumulative odds ratio. Alternatively one could use the adjacent-category logit where $\exp(\beta)$ would denote the common odds ratio for all adjacent pairs of responses. Note that as the number of centers increase, the number of parameters in model (5.1) increases as well. Thus we see that the fixed effects approach is useful only when the number of centers is not too large, relative to the overall sample size.

If one suspects that the treatment effect may vary across centers, one can generalize model (5.1) to allow a separate association parameter β_i for each center. This heterogeneous association model has the form

$$\eta_{ijr} = \alpha_r + \gamma_i + \beta_i x_j, \quad r = 1, \cdots, q. \tag{5.2}$$

Model (5.2) assumes common threshold parameters across all centers, but allows for varying association parameters. A disadvantage of this model is that one does not obtain a single measure for describing the treatment effect. We will see in Section 5.3 that the random effects approach to model (5.2) provides a mean treatment effect along with an estimate of its variability.

5.2.2 Mantel-Haenszel Approach

As an alternative to the maximum likelihood estimate of β from model (5.1), Liu and Agresti (1996) proposed a Mantel-Haenszel-type estimator for estimating a common cumulative odds ratio for several stratified $2 \times R$ tables. For $R = 2$, the Mantel-Haenszel estimator (Mantel and Haenszel 1959) for a common odds ratio is consistent in both large and sparse sample asymptotics. That is, when the number of centers are fixed and the sample size within each center become large, and when the number of centers increases proportionally with the overall sample size. In contrast, the odds ratio estimate obtained from model (5.1) is inconsistent under sparse asymptotics. Liu and Agresti (1996) showed that the Mantel-Haenszel-type estimator maintained the asymptotic behavior of the Mantel-Haenszel estimator and had little efficiency loss compared to the maximum likelihood estimator $\exp(\hat{\beta})$ from model (5.1) when the data was not sparse.

For the ith center with multinomial samples $\mathbf{y}_{i1}$ and $\mathbf{y}_{i1}$ and multinomial sample sizes n_{i1} and n_{i2}, respectively, the Mantel-Haenszel estimator of the common cumulative odds ratio is

$$\mathrm{OR}_{MH} = \frac{\sum_{i=1}^{n}\sum_{k=1}^{q} y^*_{i1k}(n_{i2} - y^*_{i2k})/n_{i\cdot}}{\sum_{i=1}^{n}\sum_{k=1}^{q} y^*_{i2k}(n_{i1} - y^*_{i1k})/n_{i\cdot}}, \tag{5.3}$$

where $y^*_{ijk} = y_{ij1} + \cdots + y_{ijk}$ and $n_{i\cdot} = n_{i1} + n_{i2}$. Liu and Agresti (1996) noted that even when the common cumulative odds ratio assumption does not hold, the estimator (5.3) provides a useful summary if the heterogeneity in the center-specific odds ratios is not large and the center-specific odd ratios are in the same direction. They also provide standard error estimates of (5.3) for when the assumption of a common cumulative odds ratio holds and for when it does not hold.

The Mantel-Haenszel estimator (5.3) of the common cumulative odds ratio provides a computationally simple alternative to the corresponding maximum likelihood estimator, one that is also consistent under sparse asymptotics. Liu and Agresti (1996) recommended the use of (5.3) over the maximum likelihood estimator when the sample sizes for most centers is five or less. In practice, however, the assumption of a common association across all centers is unrealistic. Thus methods for investigating and describing such heterogeneity are important. In the next section we utilize random effects to allow for heterogeneity in the cumulative odds ratios and to describe the magnitude of variability.

5.3 Random Effects Approach

Due to the recent advances in computing power, greater availability of software, and the plethora of recent literature on the topic, random effects modeling is being utilized in a wide variety of experimental situations that have clustered or longitudinal data. If one assumes that centers in a multi-center clinical trial are randomly selected from a population of centers, then a random effects approach can be used to account for heterogeneity in the centers and in the associations across centers. This approach has been examined by a number of authors when the response from the clinical trial is binary (see, e.g., Agresti and Hartzel 1999). Alternative approaches exist for incorporating heterogeneity (see, e.g., Skene and Wakefield 1990), however we will concentrate on the random effects approach.

As the research on random effects models for ordinal data has generally lagged behind that for binary data, there has been little work in applying such models to the clinical trial setting. Lindsey et al. (1997) considered the analysis of a seasonal rhinitis (or, more commonly, hay fever) clinical trial. Subjects were observed for 28 days for which they recorded a symptom response scored on a 0-3 scale, with 0 being no symptom and 3 being bad. The structure of the data differs from the multi-center data given in Table 5.1, but does exhibit clustering at the subject level. Lindsey et al. (1997) proposed a continuation-ratio logit model that conditioned on the previous subject's response, yielding a form of Markov chain. Such an approach can account for the serial correlation in time for a subject, but does not account for heterogeneity among the subjects. A random effects approach (as discussed in Chapter 3) could be used to account for such heterogeneity. A complicating factor in the dataset, however, was that it consisted of 416 patients with approximately 10,650 total observations. Thus, a random effects approach, which must approximate integrals for each subject, would certainly be computationally burdensome to fit.

In this section we consider random effects models for modeling ordinal responses from a multi-center clinical trial. We begin, in Section 5.3.1, by considering the simplest random effects model which allows a shifting in thresholds for the centers. This model assumes that a common association parameter holds for all centers. We then consider a more realistic model in Section 3.5.2 that allows the association parameter to vary as well. This heterogeneous association model is computationally more complicated as it includes a random effect for center as well as a random center-by-treatment interaction. For both models, parametric and nonparametric assumptions can be made concerning the distribution of the random effects. For this latter assumption, we discuss in Section 3.5.3 how one can extend the NPML approach of Chapter 4 to fit the heterogeneous association model, as well as general multiple random effects models.

5.3.1 Homogeneous Association

In the previous section we considered the fixed effects model (5.1) that included a separate effect for each center and a common association parameter across all centers. When the number of centers becomes large or the data become sparse within each center, problems can occur in the estimation of model (5.1). Indeed, for binary data, Agresti and Hartzel (1999) showed that infinite estimates (in absolute value) for the center effects can occur in the latter situation. As an alternative to model (5.1), one can treat the centers as if they were a random sample from a population of centers. This amounts to allowing for a shifting of thresholds by center. Thus the linear predictor is given by

$$\eta_{ijr} = \alpha_r + \beta x_j + u_i, \quad r = 1, \cdots, q, \tag{5.4}$$

where u_i denotes the random effect for the ith center, $i = 1, \cdots, n$, and the remaining parameters and covariate are defined as in Section 5.2.1.

The specification of the homogeneous random effects model is completed by specifying the distribution of the random effect u_i and by choosing a link function. As discussed in Chapters 3 and 4, one can make a parametric or nonparametric assumption about the distribution of the random effect. From our experience and from the results of the simulation study in Section 4.5.1, both approaches will yield similar estimates for the association parameter. Regardless of the choice, the algorithms given in Chapters 3 and 4 can be used to obtain maximum likelihood estimates of the parameters. The choice of the link function will depend on the form of the ordinal response and the desired interpretation. For example, if the responses have a sequential ordering, then the continuation-ratio logit link would be most appropriate. For the data given in Table 5.1, the cumulative logit or adjacent-category logit link would most likely be chosen.

An advantage of using model (5.4) over model (5.1) is that non-infinite estimates of the center effects can be obtained, even when the data are sparse within the centers. As discussed in Section 3.4.3, estimates of the center effects in model (5.4) are obtained by calculating the expected values of the $\{u_i\}$ given the data and the final parameter estimates. These predictions are analogs of best linear unbiased predictors (BLUP) for mixed models with normal responses. As the prediction for a given center effect is a function of the parameter estimates obtained from all centers, the predictions "borrow" information from all centers. Thus non-infinite estimates can be obtained for centers for which $\hat{\gamma}_i$ would be infinite for maximum likelihood fitting of model (5.1). Center estimates can be obtained for both the parametric and nonparametric distributional assumptions. However, due to the discreteness of the random effects distribution, the set of possible center estimates for the nonparametric approach is smaller than that of the parametric approach. Thus, the parametric approach is more well suited for obtaining such predictions.

Model (5.4) assumes that the association parameter β is the same across all centers. It also assumes that the distance between thresholds is the same for all centers. For the continuation-ratio and adjacent-category logit links, one could generalize (5.4) by replacing u_i with u_{ir}, allowing the thresholds to vary individually. For the cumulative logit link, the extended model of Tutz and Hennevogl (1996) (see Section 3.7) could also be applied. For all three cases, one no longer has a single estimate for the center effects. Indeed, for the extended cumulative logit model, the predicted threshold effects would have little meaning as they do not correspond to the original thresholds (expect for the first threshold). In general, such models would most likely provided similar estimates of β as provided by model (5.4). A more beneficial generalization would be to allow the association parameter to vary over centers. We consider this heterogeneous association model in the next section.

5.3.2 Heterogeneous Association

The assumption of a common association parameter for all centers is unrealistic. Indeed one would expect the parameter to vary at least nominally due to variation in, for example, equipment, personnel, or patients from center to center. One can extend the homogeneous random effects model to allow for a varying association parameter by incorporating a center-by-treatment interaction into model (5.4). Since the center effects are already assumed to be random, the resulting interaction is also random. The heterogeneous random effects model can be written in the form

$$\eta_{ijr} = \alpha_r + \beta x_j + u_i + v_{ij}, \quad r = 1, \cdots, q, \quad j = 1, 2, \tag{5.5}$$

where the complete random effects vector for the ith subject is $\mathbf{u}_i' = (u_i, v_{i1}, v_{i2})$, $i = 1, \cdots, n$. We assume that $\mathbf{u}_i$ follows a distribution G with mean $\mathbf{0}$ and covariance matrix Σ.

The form of the heterogeneous random effects model (5.5) has been considered previously by a number of authors. Littell et al. (1996) used this model to analyze data from an eight-center clinical trial in which two drugs were compared with respect to binary response. Likewise, Booth and Hobert (1999) analyzed data from fourteen retrospective studies on the association between smoking and lung cancer. Such data have the form of Table 5.1 with center numbers replaced by study numbers. As noted by Booth and Hobert (1999), model (5.5) has three random effects, which, under the assumption of normality for the random effects, means that three integrals must be approximated for each center. In fact, a simple reparameterization of the random effects in (5.5) can reduce the number of intractable integrals to two for each center. This simplification was used by Agresti and Hartzel (1999) for analyzing binary multi-center data.

Reparameterization

Let the random effects covariance matrix Σ for model (5.5) have the unstructured form

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 \end{bmatrix}. \tag{5.6}$$

Model (5.5) can be simplified in the following manner

$$\begin{aligned} \eta_{ijr} =& \alpha_r + \beta x_j + u_i + v_{ij} \\ =& \alpha_r + \beta x_j + u_i + \frac{v_{i1} + v_{i2}}{2} - \frac{v_{i1} + v_{i2}}{2} + v_{ij} \\ =& \alpha_r + \beta x_j + u_i^* + (-1)^{I[j=1]} v_i^*, \end{aligned} \tag{5.7}$$

where $u_i^* = u_i + \dfrac{v_{i1} + v_{i2}}{2}$, $v_i^* = \dfrac{v_{i1} - v_{i2}}{2}$, and $I[j = 1]$ is an indicator function which is one when $j = 1$ and zero otherwise. In model (5.7) there are only two random effects $\mathbf{u}_i' = (u_i^*, v_i^*)$ and only two integrals to approximate. The new covariance matrix Σ^* is of the form

$$\Sigma^* = \begin{bmatrix} \sigma_1^{2*} & \sigma_{12}^* \\ \sigma_{12}^* & \sigma_2^{2*} \end{bmatrix}. \tag{5.8}$$

It is informative to consider the relationship between the elements in (5.8) and the elements in (5.6). Using the definitions of u_i^* and v_i^*, one can show that

$$\begin{aligned} \sigma_1^{2*} =& \sigma_1^2 + \frac{1}{4}(\sigma_2^2 + \sigma_3^2) + \sigma_{12} + \frac{\sigma_{23}}{2}, \\ \sigma_2^{2*} =& \frac{1}{4}(\sigma_2^2 + \sigma_3^2) - \frac{\sigma_{23}}{2}, \\ \sigma_{12}^* =& \frac{1}{2}(\sigma_{23} - \sigma_{13}) + \frac{1}{4}(\sigma_2^2 - \sigma_3^2). \end{aligned}$$

If, for example, one assumed in model (5.5) that the interaction components v_{i1} and v_{i2} were normally distributed and had common variances and covariances (i.e. $\sigma_2^2 = \sigma_3^2$

and $\sigma_{12} = \sigma_{13}$), it would translate into assuming that u_i^* and v_i^* are independent in model (5.7) (i.e. $\sigma_{12}^* = 0$). Since model (5.7) requires fewer integrals to be approximated, it is advantageous to use that representation. Thus, for the remainder of the chapter when referring to the heterogeneous association model, we will be referring to model (5.7). Regardless of the model fit, if one allows both (5.6) and (5.8) to be unstructured, both models will yield the same estimates and log-likelihood values.

Estimation

Estimation of the heterogeneous association model (5.7), under the assumption of normality for the random effects, can be carried out using any of the algorithms discussed in Chapter 3. Recall that the general multinomial random effects model was of the form $\boldsymbol{\eta}_{ij} = Z_{ij}\boldsymbol{\beta} + W_{ij}\mathbf{u}_i$. For model (5.7), Z_{ij} consists of the appropriate design matrix for the chosen response function, and a column consisting of a 1.0 or 0.0 depending on the status of the treatment. The corresponding parameter vector is $\boldsymbol{\beta}' = (\alpha_1, \cdots, \alpha_q, \beta)$, and the random effects vector is $\mathbf{u}_i' = (u_i^*, v_i^*)$. The random effects design matrix W_{ij} has a column of ones for the random center effect and a column consisting of -1.0 or 1.0 depending on the status of the treatment.

In addition to estimating the average association β across all centers and its variability σ_2^{2*}, one can also obtain predictions for the value of the association parameter for each center. This is analogous to the predicted center effects discussed in the previous section for model (5.4). To accomplish this, one would predict the value of v_i^* in model (5.7) for each center, and then use the estimated value of β to determine the association for each center. If one were using a cumulative logit link for example, one could exponentiate these predictions to obtain the center-specific cumulative odds ratios.

For data arising from a multi-center clinical trial, the number of centers is usually small. Thus it is questionable whether valid estimates can be obtained for the covariance matrix Σ^*. Indeed for Table 5.1 the estimates would be based on only eight observations (centers). In Section 5.4 we report the results of a simulation study in which we examined the performance of the heterogeneous association model under the assumption of normality for the random effects. As an alternative to normality, one could also estimate model (5.7) using a nonparametric approach. In the next section we show how one can extend the NPML model of Chapter 4 to fit the heterogeneous association model.

5.3.3 Heterogeneous Association: NPML Estimation

In Section 4.2 we outlined the NPML EM algorithm for fitting multinomial random effects models with shifted thresholds. With only slight modifications, this algorithm can also be used to fit model (5.7), or, more generally, models with multiple random effects. There have been relatively few uses of the NPML approach with multiple random effects in the literature. Davies and Pickles (1987) utilized a bivariate discrete random effects distribution for studying shopping travel. For their particular model, the NPML estimate of the mixing distribution had a surprisingly large estimated support size of eighteen. Davies (1993) applied the NPML approach to the modeling of residual heterogeneity in recurrent behavior data. As an example, he considered a depression dataset in which the depression status of subjects (depressed or not depressed) was recorded at each of four consecutive interviews. Using a first-order Markov chain model, Davies (1993) assumed that the transition matrix consisted of two subject-specific transition probabilities which he modeled with a bivariate discrete distribution. For his application, the NPML estimate required only three mass points. Finally, Aitkin (1999) discussed how one could fit random coefficient models using the NPML approach. He utilized this approach to analyze a binary multi-center clinical trial dataset allowing for heterogeneous associations. Though the estimated discrete bivariate distribution had only three mass points, Aitkin (1999) noted that, in general, more mass points were needed for the bivariate case. From our experience, only three to five mass points are typically needed for the bivariate case as well.

For a model such as (5.7), the NPML approach assumes that the joint distribution of (u_i^*, v_i^*) is a discrete distribution with mass points (m_{k1}, m_{k2}) and masses p_k, $k = 1, \cdots, K$, where K is the unknown support size. For the more general model, the likelihood can be written in the form

$$L(\boldsymbol{\beta}, \mathbf{p}, \mathbf{m}) = \prod_{i=1}^{n} \sum_{k=1}^{K} p_k \, f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \mathbf{m}_k), \tag{5.9}$$

where $f(\tilde{\mathbf{y}}_i \mid \boldsymbol{\beta}, \mathbf{m}_k) = \prod_{j=1}^{T_i} f(\tilde{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}, \mathbf{m}_k)$, $\mathbf{m}_k' = (m_{k1}, \cdots, m_{km^*})$, and m^* is the dimension of the random effects. Likelihood (5.9) is a generalization of likelihood (4.2) given in Section 4.2.

Maximization of the log of (5.9) can be accomplished using the NPML EM algorithm defined in Section 4.2. For model (5.7) we now have K pairs of mass points (m_{k1}, m_{k2}) to estimate. We use a similar approach for incorporating the pairs of mass points into the design matrix Z_{ij} as was used to incorporate the single mass point m_k for the shifted threshold model. Specifically, in addition to the K level factor used in the shifted threshold model, we also include a second factor that is obtained by interacting the first factor with the covariate x_j. Recall that the K level factor is coded as a set of K dummy variables. Thus the covariate x_j is multiplied by each of the K dummy variables. To avoid identifiability issues with the thresholds, one must exclude two of the design columns related to the mass point factors.

In a similar manner, one can use the NPML approach for fitting models, such as the continuation-ratio logit or baseline-category logit, that allow for varying thresholds. Recall that in these models, one specifies separate regression parameters for each logit. To allow for varying thresholds, one simply incorporates a K level factor for each logit. To identify all parameters, one must either suppress the intercept parameter for each logit, or remove one of the dummy variables for the K level mass point factor. For the toxicity and life satisfaction datasets considered in Section 3.6, we used a parametric random effects approach to fit a continuation-ratio logit model and a baseline-category logit model that allowed for varying thresholds. By assuming a bivariate discrete distribution for the random effects, we can now use a nonparametric random effects approach to fit these same models. To this end, we briefly reconsider the toxicity and life satisfaction datasets originally analyzed in Section 3.6.

Developmental toxicity data

In Section 3.6.2 we analyzed data from a toxicity study (Table 3.3), in which pregnant mice were administered one of four dosages (0, 0.75, 1.50, 3.00 g/kg) of ethylene glycol. After exposure, their fetuses were then examined for defects, and classified as either Dead/Resorption, Malformed, or Normal. Results from Table 3.9 suggested that a continuation-ratio logit model allowing separate dosage parameters for the two logits, which model the probability of a dead/resorbed fetus and the conditional probability of a malformed fetus given the fetus was alive, as well as random threshold parameters was adequate. Thus, the linear predictor for the ith litter and the rth logit is

$$\eta_{ir} = \alpha_r + \beta_{DO_r}\, x_i + u_{ir}, \tag{5.10}$$

where $\mathbf{u}_i = (u_{i1}, u_{i2})'$ is a bivariate random effect, and the remaining parameters and covariate are defined as in Section 3.6.2.

Previously we assumed that $\mathbf{u}_i$ followed a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix Σ. We now relax this assumption and assume that $\mathbf{u}_i$ follows a bivariate discrete distribution with mass points (m_{k1}, m_{k2}) and masses p_k, $k = 1, \cdots, K$, where K is the unknown support size. Table 5.2 contains the NPML estimate for model (5.10) along with the results from the corresponding parametric model which allowed for correlation between thresholds. Assuming a discrete joint distribution for u_{i1} and u_{i2} automatically incorporates correlation between the random effects. In fact, we were unable to fit the NPML model which assumes that the random effects have zero correlation.

The estimated discrete distribution contained three points with locations (-3.74, -3.31), (-19.67, -5.31), and (-3.87, -7.76), and masses 0.44, 0.22, and 0.33, respectively. Note the extremely small first coordinate (-19.67) for the second mass point, which pertains to the random effect for the first logit. Observing this coordinate during the estimation algorithm revealed that it was decreasing towards $-\infty$ with little to no change in the parameter and log-likelihood values. A possible reason for this behavior is that 82% of the litters contained no dead or reabsorbed fetuses, with the maximum number of dead or reabsorbed fetuses being two in the remaining litters. As a result of this small coordinate, the estimate of the standard deviation for the first threshold and the correlation estimate are suspect, as are the estimates for the threshold parameters as they are functions of the mass points. Examining the dosage parameters we see that similar results are found for both the parametric and nonparametric approaches, though the NPML approach does have a larger estimate and standard error for the dosage parameter in the second logit. Conclusions based on the dosage parameters, however, would remain the same for both approaches.

Life satisfaction data

In a similar manner, we now can nonparametrically analyze the item response data (Table 3.4) originally analyzed parametrically in Section 3.6.3. Recall that subjects were asked to rate their degree of life satisfaction with their family, hobbies, and residence using a three-point scale (1 = Low, 2 = Medium, 3 = High). In Section 3.6.3 we utilized a baseline-category logit model that allowed for correlated random threshold effects in order to estimate the item parameters. This model was of the form

$$\eta_{ijr} = \beta_{F_r}\ x_{ij1} + \beta_{H_r}\ x_{ij2} + \beta_{R_r}\ x_{ij3} + u_{ir}, \tag{5.11}$$

Table 5.2: Parameter estimates for model (5.10) from both the NPML algorithm and the adaptive Gauss-Hermite algorithm (AGH) with 18 quadrature points using the continuation-ratio logit link.

	NPML	AGH
α_1	-7.284	-4.198
α_2	-5.242	-4.356
β_{DO_1}	0.092	0.083
	(.211)	(.217)
β_{DO_2}	2.320	1.780
	(.276)	(.219)
σ_1	6.567	0.559
σ_2	1.948	1.587
ρ	0.027	0.080
LL	-459.532	-464.733

where u_{ir} is a threshold-specific random effect, and the remaining parameters and covariate are defined as in Section 3.6.3.

Using an assumption of normality for the random threshold effects that allowed for correlation, we estimated the item parameters for each of the two logits. Alternatively one could assume that the threshold random effects follow an unspecified discrete bivariate distribution, and use the NPML algorithm to estimate the item parameters. Table 5.3 contains the results for this model, as well as the corresponding parametric results using the adaptive Gauss-Hermite algorithm (AGH) with 15 quadrature points. As we did with the adaptive algorithm, we modified the NPML algorithm to exploit the fact that only 27 unique response profiles were possible for this experiment.

The NPML estimate of the discrete distribution had a support size of four. The mass points and masses were given by $\{(4.39, 2.74),\ (1.55, 0.64),\ (0.08, 1.11),\ (-2.71, -1.33)\}$ and $\{(.22, .40, .35, .03)\}$, respectively. The results for both models were very similar, including the estimates for the standard deviations of the random threshold effects. One might consider using the NPML approach over the adaptive

Table 5.3: Parameter estimates for model (5.11) from both the NPML algorithm and the adaptive Gauss-Hermite algorithm (AGH) with 15 quadrature points using the baseline-category logit link.

	NPML	AGH
β_{F_1}	1.466	1.384
	(.156)	(.169)
β_{H_1}	0.997	0.933
	(.122)	(.119)
β_{R_1}	1.209	1.144
	(.119)	(.115)
β_{F_2}	3.300	3.264
	(.160)	(.161)
β_{H_2}	1.732	1.709
	(.120)	(.118)
β_{R_2}	1.530	1.509
	(.122)	(.118)
σ_1	0.898	0.832
σ_2	1.723	1.626
ρ	0.862	0.617
LL	-3734.85	-3736.93

approach purely for computational reasons, as the latter requires numerical approximation of two integrals. However, the computational effort for the NPML approach with multiple random effects is relatively high as well. Matrices in the algorithm quickly become large as an additional K dummy variables are needed for each additional random effect. Use of the acceleration techniques discussed in Section 4.2 is highly recommended when the number of random effects is increased.

5.4 Simulation Study

In the previous section we showed how one could use the parametric approaches of Chapter 3 to analyze multi-center clinical trial data. If one is willing to assume that the centers constitute a random sample from some population of centers, one can estimate the variability among the centers as well as the variability in the association parameter across the centers. As the estimation algorithms in Chapter 3 will provide estimates for the heterogeneous association model for the majority of datasets, it is easy to blindly apply such an approach and interpret the resulting estimates. However, consider the data in Table 5.1. In this multi-center clinical trial, data were collected at only eight centers. Suppose we actually observed the random center and interaction effects for each center. Thus we would have a sample of eight pairs of realizations that we would assume, under the parametric homogeneous association model, arose from a bivariate normal distribution with zero mean and unknown covariance matrix. The standard multivariate sample variance formula could then be used to estimate the unknown covariance matrix. Indeed, we would probably have little faith in this estimate as it is based on only eight observations. When fitting the heterogeneous association model, we do not have the realizations of the random effects and must use the information in the data to learn about their behavior. Thus, it is questionable how well we can describe their distribution based on such small numbers of centers.

In order to assess the performance of the adaptive algorithm for fitting the parametric heterogeneous association model to multi-center clinical trial data, we performed a number of simulations. For the heterogeneous association model (5.7), there are a large number of factors that could be examined in a simulation study. For example, the number of centers, the sample size within centers, the distribution of the random effects, the structure of the covariance matrix for the random effect, etc. To study all possible factors would be impractical in a single simulation study. In addition, fitting model (5.7) requires the approximation of two integrals for each center. Thus each simulation entails a high amount of computation and time. We also found for datasets such as Table 5.1 with small numbers of clusters, that a large simulation size is required to reduce the Monte Carlo error to a reasonable level.

In light of these remarks, we performed a series of simulations using the heterogeneous random effects model (5.7)

$$\eta_{ijr} = \alpha_r + \beta x_j + u_i^* + (-1)^{I[j=1]} v_i^*, \tag{5.12}$$
$$r = 1, \cdots, q = R-1, \quad i = 1, \cdots, n, \quad j = 1, 2,$$

where we set the number of responses $R = 3$, $\alpha_1 = -1$, $\alpha_2 = .5$, and $\beta = .75$. We also let $x_j = 1$ if $j = 2$ and zero otherwise. We considered two sample size structures for the data. In the first sample size structure, we used a small number of centers (8), but a large number of observations for each treatment within center (30). For the second sample size structure, we increased the number of clusters to 20, but reduced the number of observations per treatment to 12. The next factor that we set was the covariance structure for the random effects. For simplicity we set the covariance term in Σ to be zero and looked at two sets of values for the variance components (σ_1^2, σ_2^2), where the first variance component corresponds to the random center effect and the second to the random interaction effect. The values considered were (0.3, 1.0) and (1.0, 0.3), which examined one situation where the random interaction component was large and the other where the random center component was large. We also looked at a third case were a random center effect existed, but the random interaction did not. For this case $(\sigma_1^2, \sigma_2^2) = (0.3, 0)$. Finally, we considered two cases for the distribution of the random effects. In the first case we assumed a bivariate normal distribution. For the second case we used a mixture of two bivariate normal distributions to produce a non-normal distribution with the required mean and covariance structures. We discuss how this was done below.

Simulating data from multivariate non-normal distributions is difficult as most do not have the flexible parameterizations for the covariance matrix as the multivariate normal distribution has. An alternative approach for simulating non-normal multivariate distributions with known covariance structures is to use a mixture of two multivariate normal distributions. In this approach, one samples from the first multivariate normal with probability p and the second with probability $1 - p$. Suppose the random variable $\mathbf{x}$ is distributed as

$$f(\mathbf{x}) = p\, MVN(\boldsymbol{\mu}_1, \Sigma_1) + (1-p)\, MVN(\boldsymbol{\mu}_2.\Sigma_2), \tag{5.13}$$

where MVN denotes the multivariate normal distribution and $0 \leq p \leq 1$. Johnson (1987, p. 57) showed that the mean and variance of $\mathbf{x}$ distributed as (5.13) are

$$\begin{aligned} E(\mathbf{x}) =& p\, \boldsymbol{\mu}_1 + (1-p)\, \boldsymbol{\mu}_2, \\ Cov(\mathbf{x}) =& p\, \Sigma_1 + (1-p)\, \Sigma_2 + p(1-p)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'. \end{aligned}$$

Using these formulas we chose values for Σ_1, Σ_2, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and p that would produce the covariance structures defined above. The values chosen were $p = 0.5$,

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 1/6 \\ 1/6 \end{bmatrix} \qquad \boldsymbol{\mu}_2 = \begin{bmatrix} -1/6 \\ -1/6 \end{bmatrix}$$

$$\Sigma_1 = \begin{bmatrix} 23/18 & -1/6 \\ -1/6 & 2/45 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 2/3 & -1/9 \\ -1/9 & 1/2 \end{bmatrix}.$$

Figure 5.1 displays a contour plot of the mixture distribution produced by the choices given above.

In summary, the factors that were selected for the simulation runs are given in the table below.

Sampling Structure	Covariance Structure	Random Effects Distribution
(# Centers, # Obs/Trt)	(σ_1^2, σ_2^2)	(bivariate)
(8, 30)	(.3, 1.0)	Normal
(20, 12)	(1.0, .3)	Mixture
	(.3, 0)	

Since the covariance structure without the random interaction was only run with a normal center effect, there were a total of ten simulations run. Pilot simulation studies revealed that the Monte Carlo error associated with the association parameter was quite large for the eight center case. Thus simulation sizes of 1,000 were used for the runs with eight centers, and sizes of 750 were used for the 20 center cases. This reduced the Monte Carlo error in the parameter estimates to approximately 0.01 for both center sizes.

Tables 5.4 and 5.5 contain the results for eight of the ten simulations performed. Missing are the two simulations in which we simulated data from a univariate normal distribution such that the random interaction did not exist. For both of these simulations (ie. cluster sizes of 8 and 20), the adaptive quadrature algorithm had severe convergence problems. In better than 85% of the simulated datasets the algorithm failed to converge. It was obvious that the cause of the convergence problem was the near zero estimate for the interaction variance component. In hope of alleviating this problem, we modified the adaptive algorithm so that the elements of the Cholesky square root of the covariance matrix were estimated instead of the actual variance components. As noted before, this approach often performs better when the variance components are small. For the bivariate random effects case, the Cholesky square

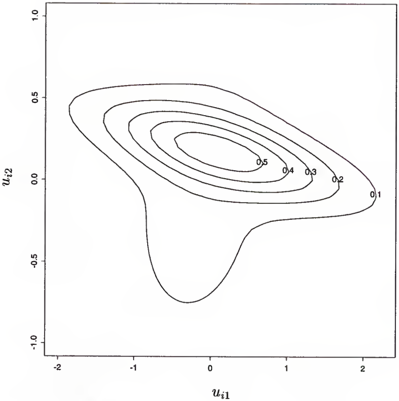

Figure 5.1: Contour plot of the bivariate normal mixture distribution used for Table 5.5.

root Q is given by

$$\begin{aligned}\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} &= Q\,Q' \\ &= \begin{bmatrix} q_{11} & 0 \\ q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} q_{11} & q_{21} \\ 0 & q_{22} \end{bmatrix}.\end{aligned}$$

Thus $\sigma_1^2 = q_{11}^2$, $\sigma_{12} = q_{11}q_{21}$, and $\sigma_2^2 = q_{21}^2 + q_{22}^2$. Unfortunately this did not remedy the convergence problems as the estimate of q_{22} was very near zero as well. We do not feel this is a problem of the algorithm itself. Indeed, if one were questioning whether to fit the heterogeneous random effects model, one can use the failure of the algorithm as a indication that it is not needed. By observing the parameter estimates at each iteration of the algorithm, one can easily see which variance component is near zero. For those simulations that did converge, the estimate for the interaction component was very near zero. In such instances, evaluation of the observed information became unstable and often times led to negative variance estimates for the parameters. Due to the unstable estimates and poor convergence rates, we excluded any tables summarizing these results.

Table 5.4 contains the results for the simulations in which the random effects were generated from a bivariate normal distribution. Reported are the estimated biases in the parameter estimates, along with the average standard error estimate of $\hat{\beta}$ calculated from the observed information matrix and the corresponding Monte Carlo estimate. Examining the bias in β, the parameter of interest, we see that the degree of bias is minor. The largest absolute estimated bias (-0.018) occurred with a cluster size of 20 and an interaction variance component of 1.0. This corresponds to a percent bias of -2.4%. As one would expect, both cluster sizes had larger estimated biases in β when the random interaction component was larger. We also see that the smaller cluster size had smaller estimated biases for β than the larger cluster size. This is due to the difference in treatment sample size between the two cluster sizes (from 30 to 12). Examining the threshold estimates, we see for the cluster size of eight that the biases nearly doubled when the variance component for the center effect was changed from .3 to 1.0. In contrast, the biases remained the same or decreased for the cluster size of 20. These results indicate that, with a larger cluster size, the heterogeneity model can more accurately estimate the threshold parameters even with a large center variance component.

We can also see some clear patterns from the biases in the variance components. First, the estimated biases for all variance components are smaller when the largest variation comes from the treatment-by-center interaction. In these simulations, the estimated bias in the center variance component was approximately -.01 (or -4% bias), regardless of cluster size. The interaction variance component exhibited greater bias, with the eight cluster size having a -13% bias. It is interesting that the 20 cluster size had only a -.5% bias for the interaction variance component (explainable by Monte Carlo error), but had a larger bias in the estimate for β. The larger bias in the estimate for β is probably due to the smaller treatment sample size. When the larger variance component was associated with the center effect, there was considerable bias in the variance component estimates. Indeed, for these cases the estimates of the covariance matrix are very poor. We suspect that the large variation in the center effect swamped the interaction variance component, causing difficulties in estimation of the variance components. In light of these poor estimates, the standard error estimates for β based on the observed information matrix were reasonable. For all simulations these estimates were lower than the corresponding Monte Carlo estimates. The largest differences occurred when the interaction variance component was large.

Table 5.5 contains the results for the simulations in which the random effects were simulated from the distribution given in Figure 5.1. Bias estimates for β are similar to those in Table 5.4 with the exception being the eight cluster size with (1, .3) which had an average bias of .03 or about 4%. It is also interesting for that setting that the threshold biases are quite small, even though the center variance component was large. In contrast, the eight cluster size with a center variance component of .3 had considerable more bias. It is difficult to explain this occurrence. We suspect it is a combination of the small center size and the form of the bivariate normal mixture. Only eight observations are selected from the mixture for each dataset, thus it is quite possible to select most observations from one of the bivariate normal components. The one bivariate normal component has a small variance component for center (2/3), which would result in reduced biases for the thresholds as seen in Table (5.5). For the cluster size of 20, the patterns in Table 5.4 held, with decreasing biases seen in both the threshold parameters and association parameter when moving from the (.3, 1) case to the (1, .3) case. The biases in this table are generally greater, however. We again see considerable biases in the variance components. For this table, all biases were large, regardless of the cluster size or variance component values. Surprisingly, the standard error estimates are still in close agreement under the mixture model.

From these simulations we see that one can obtain fairly accurate estimates of the association parameter and its standard error, even when the cluster size is small, or the treatment sample size is small. Estimates for the variance components can be much less accurate, especially if the random effects distribution deviates from normality. For extremely small cluster sizes or extremely sparse data, we would expect the results to continue to deteriorate. Thus caution must be taken when interpreting estimates such as the variance components. Though not examined here, one can also expect predictions of the center effects and center-specific cumulative log odds ratios to be poor as well, as they are based on the estimates of the parameters, and in particular the random effects distribution estimates.

Table 5.4: Estimated bias of parameter estimates for the parametric heterogeneous random effects model (5.12) using simulated multi-center clinical trial data with treatment sample sizes of n_{ij}. The random effects were simulated from a bivariate normal distribution with covariance structure (σ_1^2, σ_2^2), and covariance term $\sigma_{12} = 0$.

	8 CLUSTERS ($n_{ij} = 30$)		20 CLUSTERS ($n_{ij} = 12$)	
	(.3, 1)	(1, .3)	(.3, 1)	(1, .3)
α_1	0.013	0.022	0.009	0.008
α_2	0.016	0.026	0.007	0.000
β	-0.003	0.001	-0.018	-0.007
$(\mathrm{SE}(\hat{\beta})_O)$	(.673)	(.431)	(.468)	(.302)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(.733)	(.419)	(.478)	(.307)
σ_1^2	-0.013	0.590	-0.011	0.662
σ_2^2	-0.130	-0.721	-0.005	-0.697
σ_{12}	0.012	0.013	-0.007	-0.002

Table 5.5: Estimated bias of parameter estimates for the parametric heterogeneous random effects model (5.12) using simulated multi-center clinical trial data with treatment samples sizes of n_{ij}. The random effects were simulated from a mixture of two bivariate normal distributions with covariance structure (σ_1^2, σ_2^2), and covariance term $\sigma_{12} = 0$ (see Figure 5.1).

	8 CLUSTERS ($n_{ij} = 30$)		20 CLUSTERS ($n_{ij} = 12$)	
	(.3, 1)	(1, .3)	(.3, 1)	(1, .3)
α_1	0.162	-0.001	0.023	0.011
α_2	-0.125	-0.001	0.033	0.014
β	-0.003	0.029	-0.015	0.001
$(\mathrm{SE}(\hat{\beta})_O)$	(.386)	(.425)	(.285)	(.306)
$(\mathrm{SE}(\hat{\beta})_{MC})$	(.382)	(.429)	(.291)	(.304)
σ_1^2	0.911	0.578	0.941	0.673
σ_2^2	-0.741	-0.698	-0.771	-0.698
σ_{12}	0.000	-0.014	0.043	-0.005

5.5 Score Tests for a Common Association Parameter

The assumption that the association parameter is exactly the same across all centers is generally unrealistic. The heterogeneous association model (5.7) is a straightforward extension of the homogeneous model that relaxes this assumption. It typically provides larger standard error estimates for the association parameter than the homogeneous model, reflecting the uncertainty in the assumption of homogeneity. For this reason, it may be recommended over the homogeneous model even when one might expect the heterogeneity to be small. Even so, it would be desirable to have a statistical test for determining if the common association assumption holds. Certainly it is more desirable for a pharmaceutical company to know that a particular drug behaves the same at all centers, as well as for them to be able to report a single estimate of the association between the drug and the placebo. Thus, in this section we consider score tests for testing that a common association holds across all centers. In terms of the heterogeneous association model (5.7), this implies that the variance component for interaction random effect v_i^* is zero, as well as the covariance term between u_i^* and v_i^*. As noted in Section 3.4.2, testing that a variance component is zero involves a nonstandard condition in that the parameter is on the boundary of the parameter space. Thus likelihood-ratio and Wald tests are not appropriate. In this section we consider score tests for testing this hypothesis.

Before reviewing some of the recent proposals for testing for zero variance components in random effects models, we mention that tests of homogeneity exist for the fixed effects heterogeneity model (5.2) as well, treating the number of centers as fixed. The common way of testing for a homogeneous association is through the use of the likelihood-ratio test comparing model (5.2) to model (5.1). As the difference in the number of parameters between the two models is $n-1$, under the null hypothesis of no interaction this test follows a χ^2_{n-1} distribution. Recently, Uesaka (1993) proposed a measure of interaction between treatment and stratum when the response variable is ordinal and there are only two treatments. Along with the proposed measure, he provided a test of the hypothesis of no interaction. In his proposals, Uesaka (1993) assumed that the strata represented a fixed factor, such as gender.

There have been a number of proposals for testing that the variance component in a generalized linear model with a single random effect is zero. Jacqmin-Gadda and Commenges (1995), who extended the work of Liang (1987), proposed a score test for testing homogeneity among clustered data adjusting for the effects of covariates. Their test was restricted to canonical generalized linear models that, under the alternative hypothesis, included only a random intercept. More recently, Lin (1997) proposed a global score test for testing that all variance components in a generalized linear mixed model are zero. The test was derived under the assumption that the random effects were independent. Her test could easily be extended to the multivariate generalized linear models considered here, however we are interested in testing that individual variance components are zero.

Very little work has been done in the area of testing that a single variance component is zero in the presence of other random effects. In addition to her global score test, Lin (1997) proposed an individual score test, and its approximation, for testing a single variance component to be zero. As in the global test, she assumed that all of the random effects were independent. Since, even under the null hypothesis, the score statistic does not have a closed form, Lin (1997) utilized the Laplace method to approximate the intractable integrals. An advantage of this approach is that the components of the score statistic can be obtained from fitting the null hypothesis model (i.e., the model without the random effect having a zero variance) using a penalized quasi-likelihood (or pseudo-likelihood) method such as that by Breslow and Clayton (1993) or Wolfinger and O'Connell (1993). Lin (1997) reported, however, that the approximate score test had similar problems with small binomial sample sizes as the penalized quasi-likelihood estimation procedure.

As we have already generalized the pseudo-likelihood method of Wolfinger and O'Connell (1993) in Section 3.5 to multinomial random effects models, generalizing the individual score test of Lin (1997) is straightforward. Thus we begin in Section 5.5.1 by outlining the Laplace approximated score test of Lin (1997) for multinomial random effects models. As noted before, this test assumes that all random effects are independent. It also has been shown to perform poorly with small binomial samples sizes, and we suspect that similar behavior could occur with small multinomial sample sizes. Thus, in Section 5.5.2 we consider a second score test for testing that a variance component (or a subset of the variance components) is zero. In this test we allow for correlated random effects, and approximate the intractable integrals using adaptive Gauss-Hermite quadrature. A disadvantage of this approach is that verification of the null hypothesis distribution is difficult due to the quadrature approximation. Thus in Section 5.5.3 we examine the performance of the proposed test through simulation. In both Sections 5.5.1 and 5.5.2 we motivate the tests for the general multinomial random effects model.

5.5.1 Laplace Approximated Score Test

Consider the complete multinomial random effects model

$$\boldsymbol{\eta} = Z\boldsymbol{\beta} + W\mathbf{u}, \tag{5.14}$$

where $Z = [Z_{ij}]$, $W = \text{diag}(W_{ij})$ and $\mathbf{u} = [\mathbf{u}_i]$. For the Laplace approximated score test, we assume that the random effects are independent and that $\mathbf{u}_i' = (u_{i1}, \cdots, u_{im})$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\Sigma = \text{diag}(\sigma_c^2)$, $c = 1, \cdots, m$. We are interested in testing the null hypothesis

$$H_o : \sigma_c^2 = 0 \quad \text{versus} \quad H_a : \sigma_c^2 > 0. \tag{5.15}$$

In the following, we partition the complete random effects vector $\mathbf{u}$ into $\mathbf{u}' = (\mathbf{u}^{-c'}, \mathbf{u}^{c'})$ where $\mathbf{u}^{c'} = (u_{1c}, \cdots, u_{nc})$ contains the random effect corresponding to the zero variance component for all n subjects, and $\mathbf{u}^{-c}$ contains the remaining $m-1$ random effects for all n subjects. Likewise, a matrix or vector superscripted by $-c$ is formed or calculated under the null hypothesis. Thus W^{-c} denotes the complete random effects design matrix with the columns pertaining to $\mathbf{u}^c$ removed.

Following Lin (1997), we write the marginal log-likelihood of model (5.14) as

$$l(\boldsymbol{\beta}, \Sigma) = \log \int \exp\{\ l(\bar{\mathbf{y}}; \mathbf{u}^c) + \log g_{\text{MVN}}(\mathbf{u}^c; 0, \sigma_c^2)\ \}\ d\mathbf{u}^c, \tag{5.16}$$

where

$$l(\bar{\mathbf{y}}; \mathbf{u}^c) = \log \int \exp\{\ \log f(\bar{\mathbf{y}} \mid \boldsymbol{\beta}; \mathbf{u}) + \log g_{\text{MVN}}(\mathbf{u}^{-c}; \mathbf{0}, \Sigma^{-c})\ \}\ d\mathbf{u}^{-c}.$$

We proceed by calculating the score statistic $s_{\sigma_c^2}(\boldsymbol{\psi}^{-c}) = \dfrac{d\, l(\boldsymbol{\beta}, \Sigma)}{d\, \sigma_c^2}$, where $\boldsymbol{\psi}^{-c'} = (\boldsymbol{\beta}', \text{vech}(\Sigma^{-c}))$. Due to the intractable integrals in (5.16) we take a Laplace expansion of $l(\boldsymbol{\beta}, \Sigma)$ about $\sigma_c^2 = 0$. This is equivalent to expanding $l(\bar{\mathbf{y}}; \mathbf{u}^c) + \log g_{\text{MVN}}(\mathbf{u}^c; 0, \sigma_c^2)$ about the true mean of the random effect $\mathbf{u}^c = \mathbf{0}$ and then integrating term by term. A two-term Taylor expansion yields

$$\begin{aligned}\exp\{l(\bar{\mathbf{y}}; \mathbf{u}^c) + \log g_{\text{MVN}}(\mathbf{u}^c; 0, \sigma_c^2)\} = \exp\{l(\bar{\mathbf{y}}; \mathbf{0})\} \Bigg(1 + \frac{d\, l(\bar{\mathbf{y}}; \mathbf{0})}{d\, \mathbf{u}^{c'}} \mathbf{u}^c + \\ \frac{1}{2} \mathbf{u}^{c'} \left[\frac{d\, l(\bar{\mathbf{y}}; \mathbf{0})}{d\, \mathbf{u}^c} \frac{d\, l(\bar{\mathbf{y}}; \mathbf{0})}{d\, \mathbf{u}^{c'}} + \frac{d^2 l(\bar{\mathbf{y}}; \mathbf{0})}{d\, \mathbf{u}^c\, d\, \mathbf{u}^{c'}} \right] \mathbf{u}^c + \epsilon \Bigg),\end{aligned} \tag{5.17}$$

where ϵ contains third and higher terms of $\mathbf{u}^c$. Denoting the subset of the random effects design matrix associated with $\mathbf{u}^c$ by W^c, we have that

$$\frac{d\, l(\bar{\mathbf{y}}; \mathbf{u}^c)}{d\, \mathbf{u}^c} = W^{c'} \frac{d\, l(\bar{\mathbf{y}}; \mathbf{u}^c)}{d\, \boldsymbol{\eta}} \quad \text{and} \quad \frac{d^2 l(\bar{\mathbf{y}}; \mathbf{u}^c)}{d\, \mathbf{u}^c\, d\, \mathbf{u}^{c'}} = W^{c'} \frac{d^2 l(\bar{\mathbf{y}}; \mathbf{u}^c)}{d\, \boldsymbol{\eta}\, d\, \boldsymbol{\eta}'} W^c. \tag{5.18}$$

Also note that

$$\frac{d\,l(\bar{\mathbf{y}};\mathbf{0})}{d\,\boldsymbol{\eta}} = D^{-c}\,R_{\boldsymbol{\pi}}^{-c^{-1}}(\bar{\mathbf{y}} - \boldsymbol{\pi}^{-c}) \tag{5.19}$$

$$\frac{d^2 l(\bar{\mathbf{y}};\mathbf{0})}{d\,\boldsymbol{\eta}\,d\,\boldsymbol{\eta}'} = O^{-c} - D^{-c}\,R_{\boldsymbol{\pi}}^{-c^{-1}}\,D^{-c'}, \tag{5.20}$$

where $D^{-c} = \text{diag}(D_{ij}^{-c})$, $R_{\boldsymbol{\pi}}^{-c} = \text{diag}(R_{\boldsymbol{\pi}_{ij}}^{-c})$, and $O^{-c} = \text{diag}(O_{ij}^{-c})$ are calculated from the null hypothesis model (see Section 2.3 for definitions of D_{ij}, $R_{\boldsymbol{\pi}_{ij}}$, and O_{ij}).

Integrating (5.17) under the moment assumptions on $\mathbf{u}^{-c}$ and then applying to (5.16) using (5.18) – (5.20), one can obtain the following expression for the score statistic

$$s_{\sigma_c^2}(\boldsymbol{\psi}^{-c}) = \frac{1}{2}\,E\Bigg[\Big\{(\bar{\mathbf{y}} - \boldsymbol{\pi}^{-c})'\,R_{\boldsymbol{\pi}}^{-c^{-1}}\,D^{-c}\,W^{c}\,W^{c'}\,D^{-c}\,R_{\boldsymbol{\pi}}^{-c^{-1}}(\bar{\mathbf{y}} - \boldsymbol{\pi}^{-c}) - \\ \text{tr}[W^{c'}\,(O^{-c} - D^{-c}\,R_{\boldsymbol{\pi}}^{-c^{-1}}\,D^{-c'})W^{-c}]\Big\}\Big|\,\bar{\mathbf{y}}\Bigg]. \tag{5.21}$$

To evaluate (5.21), one would plug in the estimates of $\boldsymbol{\Psi}^{-c}$ obtained under the null hypothesis model. These estimates can be obtained by fitting the reduced multinomial random effects model that omits the random effect $\mathbf{u}^{c}$. For this test, we will use the restricted maximum likelihood estimates obtained from the pseudo-likelihood approach of Wolfinger and O'Connell (1993). Letting

$$F_{\sigma_c^2\sigma_c^2} = E\left(\frac{dl}{d\sigma_c^2}\,\frac{dl}{d\sigma_c^2}\right), \quad F_{\boldsymbol{\psi}^{-c}\sigma_c^2} = E\left(\frac{dl}{d\boldsymbol{\psi}^{-c}}\,\frac{dl}{d\sigma_c^2}\right), \tag{5.22}$$

$$\text{and}\quad F_{\boldsymbol{\psi}^{-c}\boldsymbol{\psi}^{-c}} = E\left(\frac{dl}{d\boldsymbol{\psi}^{-c}}\,\frac{dl}{d\boldsymbol{\psi}^{-c'}}\right),$$

where l denotes the marginal log-likelihood (5.16), the score statistic for testing (5.15) is given by

$$\lambda_s = \frac{s_{\sigma_c^2}(\widehat{\boldsymbol{\psi}}^{-c})}{\left\{F_{\sigma_c^2\sigma_c^2} - \left(F_{\widehat{\boldsymbol{\psi}}^{-c}\sigma_c^2}\right)'\left(F_{\widehat{\boldsymbol{\psi}}^{-c}\widehat{\boldsymbol{\psi}}^{-c}}\right)^{-1}\left(F_{\widehat{\boldsymbol{\psi}}^{-c}\sigma_c^2}\right)\right\}^{1/2}}. \tag{5.23}$$

Note that (5.23) is calculated under $\sigma_c^2 = 0$. Under the null hypothesis, it can be shown that λ_s follows an asymptotically standard normal distribution. Here the asymptotics refer to the number of clusters going to infinity while the number of observations on each cluster remains bounded. The proof of this for the multinomial random effects models follows directly from Lin (1997, see text and Appendix 2).

The score test in (5.23) contains intractable integrals in both the numerator and denominator. Lin (1997) used the Laplace method to approximate these sets of intractable integrals. This parallels what was done in (5.17), with $\exp\{l(\bar{\mathbf{y}};\mathbf{u}^c) + \log g_{\texttt{MVN}}(\mathbf{u}^c; \mathbf{0}, \sigma_c^2)\}$ replaced by $\exp\{l(\bar{\mathbf{y}};\mathbf{u}^{-c}) + \log g_{\texttt{MVN}}(\mathbf{u}^{-c}; \mathbf{0}, \Sigma^{-c})\}$ and the expansion taken about $\mathbf{u}^{-c} = \widehat{\mathbf{u}}^{-c}$, where $\widehat{\mathbf{u}}^{-c}$ denotes the maximum point of $l(\bar{\mathbf{y}};\mathbf{u}^{-c}) + \log g_{\texttt{MVN}}(\mathbf{u}^{-c}; \mathbf{0}, \Sigma^{-c})$.

Following her derivation exactly, it can be shown that the Laplace approximated score statistic (5.21), evaluated under the null hypothesis and at the restricted maximum likelihood estimates $\hat{\boldsymbol{\psi}}^{-c}$, is

$$s^*_{\sigma_c^2}(\hat{\boldsymbol{\psi}}^{-c}) = \frac{1}{2}\left\{(\check{\mathbf{y}}^{-c} - Z\hat{\boldsymbol{\beta}})'\, V^{-c^{-1}}\, W^c\, W^{c'}\, V^{-c^{-1}}(\check{\mathbf{y}}^{-c} - Z\hat{\boldsymbol{\beta}}) - \mathrm{tr}(W^{c'}\, P^{-c}\, W^c)\right\}, \tag{5.24}$$

where $\check{\mathbf{y}}$ and V are defined as in Section 3.5.2 with ϕ set to 1.0 and $\mathbf{R}$ equal to the identity matrix, and

$$P^{-c} = V^{-c^{-1}} - V^{-c^{-1}}\, Z\, (Z'\, V^{-c^{-1}}\, Z)^{-1} Z'\, V^{-c^{-1}}.$$

The denominator of (5.23) is then approximated by taking the expected value of the square of (5.24). Let

$$M_{ll'} = \frac{1}{2}\,\mathrm{tr}(W_{l'}'\, P^{-c}\, W_l\, W_l' P^{-c} Z_{l'}), \quad l, l' = 1, \cdots, m,$$

where W_l denotes the columns of W pertaining to the lth random effect. Then the denominator is approximated by

$$F^*_{\sigma_c^2\sigma_c^2} - \left(F^*_{\widehat{\boldsymbol{\psi}}^{-c}\sigma_c^2}\right)' \left(F^*_{\widehat{\boldsymbol{\psi}}^{-c}\widehat{\boldsymbol{\psi}}^{-c}}\right)^{-1} \left(F^*_{\widehat{\boldsymbol{\psi}}^{-c}\sigma_c^2}\right),$$

where $F^*_{\sigma_c^2\sigma_c^2} = M_{cc}$, $F^*_{\widehat{\boldsymbol{\psi}}^{-c}\sigma_c^2}$ is an $(m-1)\times 1$ vector with elements M_{lc} $(l \neq c)$, and $F^*_{\widehat{\boldsymbol{\psi}}^{-c}\widehat{\boldsymbol{\psi}}^{-c}}$ is an $(m-1)\times(m-1)$ matrix with elements $M_{ll'}$ $(l, l' \neq c)$.

An advantage of using the restricted maximum likelihood estimators $\widehat{\boldsymbol{\psi}}^{-c}$ from the pseudo-likelihood method is that a number of the components of the approximated score statistic, such as the pseudo-vector $\check{\mathbf{y}}^{-c}$, are by-products of the algorithm. Thus one can use the pseudo-likelihood algorithm defined in Section 3.5.2 to obtain some of the quantities needed to calculate the approximated score statistic. Simulation work by Lin (1997) showed that the approximated score statistic performed poorly when the binomial sample sizes were small. Similar studies could also be performed for the multinomial random effects models. We suspect that similar behavior will occur for these models as well.

5.5.2 Adaptive Gauss-Hermite Quadrature Approximated Score Test

We now consider testing that a subset of the variance components in the multinomial random effects model is zero, where the random effects are allowed to be correlated. For this test we use adaptive Gauss-Hermite quadrature to approximate the score statistic and the corresponding information matrices. For this derivation we consider the multinomial random effects model at the center level. That is,

$$\boldsymbol{\eta}_i = Z_i\boldsymbol{\beta} + W_i\mathbf{u}_i, \tag{5.25}$$

where $Z_i = [Z_{ij}]$, $W_i = [W_{ij}]$, and $\mathbf{u}_i$ is assumed to be multivariate normal with mean $\mathbf{0}$ and covariance matrix Σ. Let $\boldsymbol{\sigma} = \text{vech}(\Sigma)$ be the unique elements of the covariance matrix Σ. Using similar notation to the previous section, we are interested in testing that a subset of these element, say $\boldsymbol{\sigma}^c$ is zero. Thus we can partition $\boldsymbol{\sigma}$ as

$$\boldsymbol{\sigma} = \begin{bmatrix} \boldsymbol{\sigma}^c \\ \boldsymbol{\sigma}^{-c} \end{bmatrix}.$$

As Chant (1974) showed that the score test retains its asymptotic properties even when a parameter is on the boundary of the parameter space, a score test for testing that $H_o : \boldsymbol{\sigma}^c = 0$ is given by

$$\lambda_s = s'_{\boldsymbol{\sigma}^c} \left[F_{\boldsymbol{\sigma}^c\boldsymbol{\sigma}^c} - F_{\boldsymbol{\sigma}^c\boldsymbol{\sigma}^{-c}} F^{-1}_{\boldsymbol{\sigma}^c\boldsymbol{\sigma}^{-c}} F_{\boldsymbol{\sigma}^{-c}\boldsymbol{\sigma}^c} \right]^{-1} s_{\boldsymbol{\sigma}^c}. \tag{5.26}$$

Under a typical null hypothesis, the distribution of λ_s would be a χ^2_C distribution, with C being the number of restricted parameters in the null hypothesis. There is some question, however, if this holds for testing $\boldsymbol{\sigma}^c = 0$. Consider a model with two correlated random effects and an unstructured 2 by 2 covariance matrix. If one wishes to test that one of the variance components, say σ_2^2, is zero, then necessarily the covariance term $\sigma_{12} = 0$. Thus, for the null hypothesis $H_o : \sigma_2^2 = \sigma_{12} = 0$, one could consider σ_2^2 as the only free parameter and argue that the λ_s should be distributed χ^2_1. However, this would mean that the distribution of the score test for testing $\sigma_2^2 = 0$ when the random effects are correlated and when they are uncorrelated would be the same. The distribution is χ^2_1 for the latter model since σ_{12} is zero under both hypotheses. We were unable to find any literature in which this situation has been addressed. We feel, as the simulation study of the next section seems to suggest, that the distribution of λ_s is χ^2_C where C is the dimension of $\boldsymbol{\sigma}^c$.

In (5.26), $s_{\boldsymbol{\sigma}^c}$ is the derivative of the marginal log-likelihood for the complete model with respect to $\boldsymbol{\sigma}^c$, evaluated under the null hypothesis model maximum likelihood estimates and at $\boldsymbol{\sigma}^c = 0$. The components, $F_{..}$, of the information matrix F from the alternative hypothesis model are similarly evaluated. Unfortunately, the elements of (5.26) do not have closed form, and must be approximated. In theory, if one can approximate the intractable integrals in (5.26) accurately, then the asymptotic distribution of the test statistic should hold.

Direct calculation of (5.26) for most variance component tests is not possible. The reason is that under the null hypothesis certain elements of the covariance matrix Σ will be zero. Only for special tests (such as testing that a covariance term is zero) will this not lead to a singular matrix. To avoid singular matrices, one can use a conditional approach in which the marginal likelihood is written conditionally upon the random effects vector associated with the nonzero variances. Specifically, we partition the random effects vector for the ith center into $\mathbf{u}_i' = (\mathbf{u}_i^{c'}, \mathbf{u}_i^{-c'})$, where $\mathbf{u}_i^c$ is the random effects vector corresponding to the zero variance components, and the covariance matrix into

$$\Sigma = \begin{bmatrix} \Sigma^{c,c} & \Sigma^{c,-c} \\ \Sigma^{-c,c} & \Sigma^{-c,-c} \end{bmatrix}.$$

Also, for notational convenience, we denote the conditional log-likelihood for the ith center as

$$l_i(\bar{\mathbf{y}}_i \mid \mathbf{u}_i) = \log \prod_{j=1}^{T_i} f(\bar{\mathbf{y}}_{ij} \mid \boldsymbol{\beta}; \mathbf{u}_i).$$

Then the marginal likelihood for the ith subject can be written

$$f(\bar{\mathbf{y}}_i; \boldsymbol{\beta}, \boldsymbol{\sigma}) = \int \left\{ \int \exp(l_i(\bar{\mathbf{y}}_i \mid \mathbf{u}_i))\, g_{\mathbf{u}_i^c|\mathbf{u}_i^{-c}}(\mathbf{u}_i^c)\, d\mathbf{u}_i^c \right\} g_{\mathbf{u}_i^{-c}}(\mathbf{u}_i^{-c})\, d\mathbf{u}_i^{-c}. \qquad (5.27)$$

Since we have assumed that $\mathbf{u}_i$ has a multivariate normal distribution, the densities $g_{\mathbf{u}_i^c|\mathbf{u}_i^{-c}}$ and $g_{\mathbf{u}_i^{-c}}$ are

$$\text{MVN}[\,\Sigma^{c,-c}(\Sigma^{-c,-c})^{-1}\mathbf{u}_i^{-c},\ \Sigma^*\,] \quad \text{and} \quad \text{MVN}[\mathbf{0},\ \Sigma^{-c,-c}],$$

respectively, where $\Sigma^* = \Sigma^{c,c} - \Sigma^{c,-c}(\Sigma^{-c,-c})^{-1}\Sigma^{-c,c}$ (see, e.g., Johnson 1987, p. 50).

To avoid singular matrices under the null hypothesis, we continue by expanding $\exp(l_i(\bar{\mathbf{y}}_i \mid \mathbf{u}_i))$ in (5.27) in a Taylor series expansion about $\widehat{\mathbf{u}}_i^c = \Sigma^{c,-c}(\Sigma^{-c,-c})^{-1}\mathbf{u}_i^{-c}$ (i.e. the conditional mean of $\mathbf{u}_i^c$). That is

$$\exp(l_i(\bar{\mathbf{y}}_i \mid \mathbf{u}_i)) = \exp(l_i(\bar{\mathbf{y}}_i \mid \widehat{\mathbf{u}}_i)) \left(1 + \frac{d\, l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, \mathbf{u}_i^{c'}}(\mathbf{u}_i^c - \widehat{\mathbf{u}}_i^c) + \right.$$
$$\left. \frac{1}{2}(\mathbf{u}_i^c - \widehat{\mathbf{u}}_i^c)' \left[\frac{d\, l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, \mathbf{u}_i^c} \frac{d\, l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, \mathbf{u}_i^{c'}} + \frac{d^2 l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, \mathbf{u}_i^c\, d\, \mathbf{u}_i^{c'}}\right] (\mathbf{u}_i^c - \widehat{\mathbf{u}}_i^c) + \epsilon_i \right), \quad (5.28)$$

where $\widehat{\mathbf{u}}_i' = (\widehat{\mathbf{u}}_i^{c'}, \mathbf{u}_i^{-c'})$. By using the conditional moment assumptions for $\mathbf{u}_i^c$, the derivative formulas given in (5.18), and integrating (5.28) with respect to $g_{\mathbf{u}_i^c | \mathbf{u}_i^{-c}}$, one obtains the marginal likelihood

$$f(\bar{\mathbf{y}}_i; \boldsymbol{\beta}, \boldsymbol{\sigma}) = \int \exp(l_i(\bar{\mathbf{y}}_i \mid \widehat{\mathbf{u}}_i)) \left\{ 1 + \right.$$
$$\left. \frac{1}{2} \operatorname{tr}\left[W_i^{c'} \left(\frac{d\, l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, \boldsymbol{\eta}_i} \frac{d\, l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, \boldsymbol{\eta}_i'} + \frac{d^2 l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, \boldsymbol{\eta}_i\, d\, \boldsymbol{\eta}_i'}\right) W_i^c\, \Sigma^*\right] + \epsilon_i^* \right\} g_{\mathbf{u}_i^{-c}}(\mathbf{u}_i^{-c})\, d\, \mathbf{u}_i^{-c},$$
$$(5.29)$$

where ϵ_i^* depends on second and higher products of the variance components. Under the null hypothesis these terms are zero, and so are ignored. For the multinomial random effects models, the trace term in (5.29) can be rewritten as

$$\frac{1}{2} \sum_{i=1}^{T_i} \left[(\bar{\mathbf{y}}_{ij} - \boldsymbol{\pi}_{ij})'\; R^{-1}_{\boldsymbol{\pi}_{ij}}\; D_{ij}'\; W_{ij}^c\; \Sigma^*\; W_{ij}^{c'}\; D_{ij}\; R^{-1}_{\boldsymbol{\pi}_{ij}}\; (\bar{\mathbf{y}}_{ij} - \boldsymbol{\pi}_{ij}) \right] +$$
$$\frac{1}{2} \sum_{j=1}^{T_i} \operatorname{tr}\left[\left(\left\{ \sum_{r=1}^{q} U_{ijr}(\bar{y}_{ijr} - \pi_{ijr}) \right\} - D_{ij}\; R^{-1}_{\boldsymbol{\pi}_{ij}}\; D_{ij}' \right) W_{ij}^c\; \Sigma^*\; W_{ij}^{c'} \right], \quad (5.30)$$

where all elements are calculated using $\widehat{\mathbf{u}}_i$.

We now have a representation of the marginal likelihood that can be used in the score test (5.26) since, under the null hypothesis, the derivatives of the log of (5.29) do not contain singular covariance matrices. The score test (5.26) requires the first derivatives of (5.29) with respect to $\boldsymbol{\sigma}^c$, as well as the second derivative matrix with respect to all of the parameters $(\boldsymbol{\beta}, \boldsymbol{\sigma})$. To complicate matters, the marginal likelihood (5.29) contains a possibly multi-dimensional intractable integral. For the intractable integrals one can use Monte Carlo or quadrature techniques for obtaining approximations. We consider the latter approach here. For the derivatives one could calculate analytical derivatives or use numerical derivatives. Analytical derivatives are extremely complicated. Note that all of the elements in (5.29) and (5.30) are calculated using $\widehat{\mathbf{u}}_i$ which is made up of $\Sigma^{c,-c}(\Sigma^{-c,-c})^{-1}\mathbf{u}_i^{-c}$. This means that the elements of the covariance matrix are not localized to the multivariate normal density term, as they are in the general multinomial random effects model. So use of standard formulas for derivatives in multivariate generalized linear models is no longer possible. For this reason we utilize numerical derivatives for evaluating (5.26).

We use the following approach to calculate the score statistic (5.26). First, we obtain maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}^{-c}$ under the null hypothesis model using the adaptive Gauss-Hermite algorithm proposed in Chapter 3. We then calculate numerical first and second derivatives of the log of (5.29) with respect to $\boldsymbol{\sigma}^c$ and $(\boldsymbol{\beta}, \boldsymbol{\sigma})$, respectively, each evaluated at $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\sigma}}^{-c}$, and $\boldsymbol{\sigma}^c = \mathbf{0}$. To calculate the numerical derivatives, we must numerically approximate the integrals in (5.29). To utilize adaptive Gauss-Hermite quadrature, we first calculate the mode of the integrand in (5.29) as a function of $\widehat{\mathbf{u}}_i$, and then the curvature of the integrand, evaluated at the mode. Using these estimates we center and scale the standard nodes from Gauss-Hermite quadrature and evaluate (5.29) with the adapted nodes. This must be performed for each center. The number of quadrature points needed to approximate (5.27) will depend on the dimension of the integral under the null hypothesis. For testing the homogeneity assumption for a dataset such as Table (5.1) (where the integral is one-dimensional under the null hypothesis) we found the score test value to be accurate to approximately four decimal places using 15 quadrature points. Using the numerical derivatives we then calculate the elements in (5.26) and compare the resulting test statistic value to the appropriate χ^2 value.

We close this section with a number of comments concerning the programming aspect of the score test. As with the algorithms discussed in Chapters 3 and 4, we used the matrix programming language OX to approximate the score test and to perform the simulation studies discussed in the next section. Due to the evaluation of the marginal likelihood (5.29) (and its derivatives) at parameter values on the boundary of the parameter space, there were times when certain matrices, such as $R_{\boldsymbol{\pi}_{ij}}$, became uninvertable during the calculation of the score statistic. This mainly occurred when we were searching for the mode of the integrand of (5.29). To avoid having the maximization routine stop due to a function evaluation failure, one should check for such situations and alert the routine to move on to a new search value. We have not encountered a situation in which the final estimate of a mode yielded uninvertable matrices. Algorithms for numerical derivatives must repeatedly evaluate the function that is being differentiated. As each evaluation of the function requires n (multiple) integral approximations, calculation of the score test can require a large amount of computational time. In addition, one must find the maximum likelihood estimates under the null hypothesis model prior to calculating the score statistic. For tests in which the null hypothesis model has multiple random effects, this alone can take considerable time.

5.5.3 Simulation Study

We now examine the use of the adaptive Gauss-Hermite quadrature approximated score test for the testing of a common association parameter in the heterogeneous random effects model (5.7). We have already seen in Section 5.4 that the heterogeneous random effects model provides poor estimates of the random effects distribution when the number of centers is small. Thus, it may be overly ambitious to test that the interaction variance component and the covariance term between the center and center-by-treatment interaction are zero. We have also seen in the previous simulation study that the estimates across simulations are quite variable, requiring a large number of simulations to reduce the Monte Carlo error. Unfortunately, with the large computational time required to compute each score statistic, we do not have the luxury of performing 1,000 simulations as in Section 5.4.

To examine the performance of the score test, we studied its behavior under the null hypothesis. Under the null hypothesis, the heterogeneous random effects model (5.7) has the interaction variance component and the corresponding covariance term set to zero. Thus, we simulated from the homogeneous random effects model

$$\eta_{ijr} = \alpha_r + \beta x_j + u_i, \qquad (5.31)$$

$$r = 1, \cdots, q = R-1, \quad i = 1, \cdots, n, \quad j = 1, 2,$$

with $R = 3$, $\alpha_1 = -1.25$, $\alpha_2 = 1.25$, and $\beta = .5$. We simulated the random effect u_i from a univariate normal distribution with zero mean and variance $\sigma_1^2 = .5$. The null hypothesis for testing that a common association parameter holds is $H_o : \sigma_2^2 = \sigma_{12} = 0$. Under the null hypothesis, this test has a χ_2^2 distribution.

We ran a number of pilot studies varying the center size n and the treatment sample size n_{ij}, and it became clear that the approximated score test was performing very poorly, especially with small center sizes. To illustrate this behavior, we ran four simulations with center sizes of 8, 30, 50, and 75, each having $n_{ij} = 30$. It is obviously unrealistic that one would have a clinical trial with 50 or 75 centers, or that you would have 30 patients per treatment at each center. However, we ran the simulations at these levels to show how the score test did improve with the center size. In Table 5.6 are the rejection rates for the score test at $\alpha = 0.01$, 0.05, and 0.10. For each center size we ran 100 simulations. Ideally we should run more simulations as we have seen that the Monte Carlo error in the parameter estimates can be quite large with small to moderate center sizes. However, the computational time require to numerically approximate the first and second derivatives of the marginal log-likelihood are quite high, and thus prohibitive of large simulation studies.

Table 5.6: Rejection rates and average score test value for the adaptive Gauss-Hermite quadrature approximated score test for testing of a common association parameter in the heterogeneous association model (5.31) with $n_{ij} = 30$.

NUMBER OF	TYPE I α RATE			AVERAGE SCORE
CENTERS	0.10	0.05	0.01	TEST VALUE
8	0.24	0.21	0.12	2.275
30	0.25	0.18	0.11	3.093
50	0.18	0.14	0.04	2.403
75	0.11	0.08	0.02	1.954

We can clearly see in Table 5.31 that the score test performs very poorly when the number of centers is small to moderate. Even with a center size of 50, the type I error rate for the score test overestimated the nominal level by a fair amount. With a center size of 75, the reported type I rates were close to the nominal levels. The average test statistic values are also reported. The Monte Carlo error associated with these average test statistic values for center sizes 8, 30, 50, and 75 are approximately 1.8, 1.3, .28, .23, respectively. This illustrates the large variability present in the 8 and 30 center size runs. The large variability with the smaller center sizes did not improve when we increased the treatment size within each center. For example, with eight centers and treatment sizes of 100 the rejection rates at $\alpha = 0.10$, 0.05, and 0.01 were 0.27, 0.18, and 0.11, respectively. As mentioned before, we would expect that the approximated score test would perform poorly for small numbers of centers due to the lack of accurate information concerning the variance components. However, there are a number of other factors relating to the derivation of the test statistic that may be adversely affecting the computed value. When we examined the actual test statistic values computed for the simulations with center size 8, we found that they ranged from -78.78 to 210.61! We found similar results for the center size of 30, though there were fewer occurrences of the extreme values. Obviously, an element(s) of the score statistic was being incorrectly estimated, or more accurately, was unable to be correctly estimated.

We feel that there are three possible reasons for the failure of the score statistic in certain situations. First, the failure could be caused by the small center size in that one is more likely to obtain an odd simulated dataset. Given certain patterns of the responses across the treatments and centers, singularities or near singularities can occur in the variance matrix $R_{\boldsymbol{\pi}_{ij}}$ for the response. This would lead to a corruption of the score statistic as well. Secondly, the erroneous values could be a result of the Laplace method used in the approximation of the score statistic (see (5.28) and (5.29)). For approximating the inner integral in (5.27), we used a two-term Taylor series expansion about the conditional mode of the multivariate normal distribution. We then integrated these term with respect to the multivariate normal distribution and ignored the higher terms. Recall that these higher terms are functions of second and higher products of variance components which are zero under the null hypothesis. These terms, however, could contribute to the calculation of the test statistic when using numerical derivatives to compute the score vector and information matrix. To see this, we need to consider how numerical derivative are calculated.

Consider a function $f(\mathbf{x})$ for which the first and second derivatives at $\mathbf{x}_0$ are required. The derivative at $\mathbf{x}_0$ can be approximated by computing

$$\frac{f(\mathbf{x}_0 + \epsilon\, \boldsymbol{\iota}) - f(\mathbf{x}_0 - \epsilon\, \boldsymbol{\iota})}{2\epsilon} \simeq \left.\frac{d\, f(\mathbf{x})}{d\, \boldsymbol{\iota}'\mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_0}, \tag{5.32}$$

where $\boldsymbol{\iota}$ is a unit vector (e.g., $(1, 0, \cdots, 0)'$) with the position of the one depending on the component of $\mathbf{x}$ being differentiated. The value ϵ is a suitably chosen step length, representing a compromise between round-off error (cancelation of leading digits when subtracting nearly equal numbers) and truncation error (ignoring terms of higher order than ϵ in the approximation). Now consider the situation in the score test in which derivatives are required for (along with the other parameters) σ_2^2 and σ_{12} evaluated at zero. We argued that the higher terms ϵ_i^* in (5.29) would be zero under the null hypothesis. But from (5.32) one can see that these terms would be evaluated using numerical derivatives, albeit at some small value ϵ away from zero. It is possible that with small numbers of centers, these small changes in the numerically approximated derivative could alter the computed test statistic value.

Thirdly, the incorrect test statistic values could be caused by a general failure of the numerical derivatives. It may be that, regardless of including the higher terms of ϵ_i^*, the numerical derivatives just perform poorly for small center sizes. Ideally to check these latter two possibilities, one could include some of the higher terms in ϵ_i^* and then use analytical derivatives to compute the score test statistic. As noted before, analytical derivative are very difficult to compute for the approximated score test. Thus we first examined the use of higher terms in the Laplace method, which we discuss below. We then considered computing analytical derivatives and were successful using the software package Mathematica (Wolfram 1996). Mathematica is a symbolic mathematical package that can output derivatives of complicated functions. We used Mathematica to calculate the first and second derivatives of the marginal log-likelihood (5.29) and imported them into our Ox program. The resulting program took approximately seven times longer to run, and the code for the analytical derivatives took almost 7 MB of file space. Thus, in practice, this approach is not recommended. We discuss below, the differences found between using numerical derivatives and analytical derivatives for the approximated score test.

Consider again the expansion of $\exp(l_i(\bar{\mathbf{y}}_i \mid \mathbf{u}_i))$ in (5.27) about $\widehat{\mathbf{u}}_i^c$, the conditional mean of $\mathbf{u}_i^c$. In (5.28) we considered only the first two terms of this expansion. We now consider the inclusion of the third- and fourth-degree polynomials of the components of $\mathbf{u}_i^c$. For the test of homogeneity, $\mathbf{u}_i^c = v_i^*$ and $\mathbf{u}_i^{-c} = u_i^*$, thus the additional terms are given by

$$T_1 = \frac{1}{6}\,\frac{d^3\, l_i(\bar{\mathbf{y}};\widehat{\mathbf{u}}_i)}{d\, v_i^{*3}}\,(v_i^* - \frac{\sigma_{12}}{\sigma_1^2}u_i^*)^3 \tag{5.33}$$

$$T_2 = \frac{1}{24}\,\frac{d^4\, l_i(\bar{\mathbf{y}};\widehat{\mathbf{u}}_i)}{d\, v_i^{*4}}\,(v_i^* - \frac{\sigma_{12}}{\sigma_1^2}u_i^*)^4. \tag{5.34}$$

Note that $l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)$ is evaluated at $\widehat{\mathbf{u}}_i' = (\frac{\sigma_{12}}{\sigma_1^2} u_i^*, u_i^*)$. As before, we integrate (5.33) and (5.34) with respect to the conditional density of $v_i^* \mid u_i^*$. Since this density is univariate normal with mean $\frac{\sigma_{12}}{\sigma_1^2} u_i^*$ and variance $\sigma_2^2 - \frac{\sigma_{12}}{\sigma_1^2}$, the integrations of (5.33) and (5.34) correspond to finding the third and fourth moments of a normal random variable with mean zero and variance $\sigma_2^2 - \frac{\sigma_{12}}{\sigma_1^2}$. Thus the integral of T_1 is zero while the integral of T_2 is

$$T_2^* = \frac{1}{8} \frac{d^4\, l_i(\bar{\mathbf{y}}; \widehat{\mathbf{u}}_i)}{d\, v_i^{*4}} (\sigma_2^2 - \frac{\sigma_{12}}{\sigma_1^2})^2. \tag{5.35}$$

To see if the inclusion of the higher order terms improved the type I error rates for the previous simulations, we added T_2^* into the braced term of (5.29) and reran the simulations. The results in Table 5.7 are very similar to those reported in 5.6. That is not suggest that the inclusion of the fourth term in the Laplace expansion was unnecessary. In general, the test statistic values obtained using the two-term Laplace approximation underestimated the true test statistic value. For example, the average difference between the score test values using the four-term Laplace approximation and the two-term Laplace approximation for cluster size of 75 was 0.018 with the maximum difference being about 0.3. Note that the average test statistic value for the center size of 75 is near the expected value of the test statistic (2.0). The Monte Carlo error associated with this average is approximately .23. To better estimate the test statistic value, we ran a final simulation using a center size of 100 with 500 runs. The simulation took approximately 130 hours to run. The resulting rejection rates at $\alpha = 0.10$, 0.05, and 0.01 were 0.11, 0.06, and 0.01, respectively, with an average test statistic value of 2.07 (Monte Carlo error of .11). Thus, the approximate test does seem to perform better as the number of centers is increased.

As noted before, we also programmed, using Mathematica, the score test using analytical derivatives and the four term Laplace approximation. The resulting

Table 5.7: Rejection rates and average score test value for the adaptive Gauss-Hermite quadrature approximated score test for testing of a common association parameter in the heterogeneous association model (5.31) with $n_{ij} = 30$ and the additional fourth term of the Laplace expansion.

NUMBER OF	TYPE I α RATE			AVERAGE SCORE
CENTERS	0.10	0.05	0.01	TEST VALUE
8	0.23	0.22	0.12	0.327
30	0.25	0.18	0.11	3.453
50	0.18	0.14	0.04	2.436
75	0.11	0.08	0.02	1.972

program took extremely long to run, due to the evaluation of the complicated derivatives. Fortunately, we found that the numerical derivatives provided results within three decimal accuracy of the analytical derivatives. Thus, one can utilize the much simpler numerical derivatives for calculating the approximated score test, without sacrificing substantial accuracy.

It is evident from these simulation studies that the adaptive Gauss-Hermite approximated score test is not appropriate for the testing of a common association parameter in the heterogeneous random effects model. Given the results of the simulation study in the previous section, this is not surprising. With small to moderate numbers of centers, it is difficult to accurately estimate the covariance matrix of the random effects. However, we have seen that the association parameter, and it standard error, can be estimated with minor bias. Even though one can not accurately estimate the variability of the association parameter or test for its significance, inclusion of the additional random effect provides standard error estimates that reflect the suspected heterogeneity. Thus we still recommend that one fits the more complicated heterogeneity model. Though it performed poorly for this application, the adaptive score test seems promising for other applications in which the number of clusters or subjects is high. Rigorous proof of the distribution of the approximated score test is still required, as is more simulation work for studying its behavior under other sampling schemes.

5.6 Application

We conclude this chapter by applying the methods proposed and discussed in this chapter to Table 5.1. Recall that this table shows preliminary results from a double-blind, parallel-group clinical trial conducted at a number of centers. The purpose of the study was to compare an active drug, for treating asthma, to a placebo. At the end of the study, researchers evaluated the patient's change in condition using a three point scale (much better, better, unchanged or worse). We will utilize the cumulative logit link for modeling the ordinal response, though similar models could be fit using the adjacent-category logit link as well. We concentrate on the interpretation of the association parameter in this model, with Table 5.8 providing a summary for the majority of the models.

We begin by reporting results for the homogeneous (5.1) and heterogeneous (5.2) fixed effects model. For the cumulative logit link, the treatment effect estimate, $\hat{\beta}$, is an estimate of the cumulative log odds ratio. For the model that assumes a common association across all centers, the estimated log odds ratio is $\hat{\beta} = .93$ with a standard error of .28. Thus, for this model, one obtains a significant treatment effect. The estimated odds that the evaluation for the active drug falls below any fixed level are $\exp(.93) = 2.5$ times the estimated odds for the placebo. If one relaxes the common association assumption and fits the heterogeneous fixed effects model (5.2), the estimated center association parameters $\{\beta_i\}$ range from $\hat{\beta}_2 = -1.62$ to $\hat{\beta}_1 = 3.03$ (see Table 5.9). The likelihood-ratio statistic for comparing the heterogeneous and homogeneous fixed effects models is 24.8 on seven degrees of freedom ($P < 0.001$) giving strong evidence that the association parameter varies across centers.

For the random effects approaches we first consider the simple homogeneous model (5.4), in which only the centers are assumed random and a common association parameter is estimated for all centers. We begin by assuming that the distribution of the random center effect is normal. For this model, one obtains similar results to the fixed effects homogeneous model, with a log odds ratio estimate of $\hat{\beta} = .95$ and a standard error of .28. The estimated standard deviation of the center random effect is .60. For comparison, we also fit the homogeneous random effects model that allows for varying thresholds. In this model each of the two thresholds are assumed to be random. The estimated log odds ratio from this model is 0.93 with a standard error of .28. Thus results are substantially the same for both approaches. One could also obtain estimates for the homogeneous random effects model under the assumption of a discrete distribution for the random effect. Using the NPML algorithm of Section 4.2, the estimated log odds ratio is .94 with a standard error of .28. As has been seen before, the parametric and nonparametric approaches provide results that are in close agreement.

The heterogeneous random effects model (5.7) allows for a random interaction between the center and treatment effect, in addition to the random center effect. Thus we obtain both an expected value for the log odd ratios of the centers, as well as an estimate of its variability. For the normal random effects version of this model, the average log odds ratio estimate is 0.92 with a standard error of .53. Thus we obtain a similar estimate of the treatment effect, but a considerably larger standard error estimate. This is due to the additional variance component for the center-by-treatment interaction and reflects the uncertainty in the assumption of a common association parameter. The estimated standard deviation of the log odds ratios among the centers is 1.22. Using the proposed score tests in the previous section, we can test that the variance component for the random interaction term is zero. Using the Laplace approximate score test, which assumes that the random effects are independent and has a standard normal distribution under the null hypothesis, the score test value was 3.60 ($P < 0.01$). Using the adaptive quadrature approximated score test, assuming that the random effects are independent, the score test value was 2.38 ($P = 0.12$). For the test that allows for correlated random effects, the value was 2.62 ($P = 0.27$).

Recall, however, that the adaptive quadrature approximated score test performed poorly for small numbers of centers in the simulations of the previous section. Thus, these latter two values are suspect. In Table 5.9 we report the predicted cumulative log odds ratio estimates for the heterogeneous random effects model. These estimates are functions of the expected value of the random interaction effect given the data for each center and the final parameter estimates. For comparison, we have included the log odds ratio estimates from the heterogeneous fixed effects model as well. One can see that the estimates obtained from the random effects model are much smoother, as the estimate for each center contains information "borrowed" from all other centers.

If one relaxes the normality assumption for the random effects distribution and fits the heterogeneous random effects model nonparametrically, one obtains an average cumulative log odds ratio estimate of .98 with a standard error of .53. The NPML estimate of the mixing distribution is a four-point bivariate distribution for the random effects (u_i^*, v_i^*). The estimated mass points were (.20, .16), (-1.32, 2.91), (-.87, -.04),and (-2.28, 1.71), with corresponding probabilities (.25, .25, .37, .13). Again we see similar results for the parametric and nonparametric approaches, even for the bivariate random effects case. Table 5.9 contains the predicted cumulative log odds ratio estimates for the heterogeneous random effects model using the NPML approach. It is unclear how one should calculate standard errors for these predictions. Naive estimates could be obtained using the maximum likelihood estimates for the model parameters along with the usual formula for the variance of the interaction component given the data. However this approach does not account for plugging in the parameter estimates. One might use a similar approach to that of Booth and Hobert (1998) to account for the plug in estimates. Note that the NPML estimates seem to form four clusters. The first cluster consists of centers with very large predicted log odd ratios (centers 1 and 5). The second cluster consists of moderate sized log odds ratios and contains only center 7. The third cluster (centers 6 and 8)

Table 5.8: Estimated treatment log odds ratio and standard error for various cumulative logit models with Table 5.1.

EFFECT	CENTER	RANDOM EFFECT DISTRIBUTION	$\hat{\beta}$	STANDARD ERROR
Homogeneous	Fixed	—	.932	.278
	Random	Normal	.947	.276
		Nonparametric	.938	.282
	Varying	Normal	.931	.282
Heterogeneous	Random	Normal	.923	.526
		Nonparametric	.978	.530

consists of small predicted log odds ratios. The final cluster consists of those centers that had predicted logs odds ratios near zero (centers 2, 3, and 4). We suspect that the clustering is due to the support size of the discrete distribution being four.

In this chapter we applied the ordinal multinomial random effects models of Chapters 3 and 4 to data from multi-center clinical trials. We have seen that such models can be used to incorporate heterogeneity in both the centers and in the association parameter across centers. However, the estimates of the heterogeneity can be very poor when the cluster size is small to moderate. In these situations we have also seen that tests concerning these parameters perform poorly as well. The heterogeneous random effects model is useful for inflating the standard error of the association parameter, but one should be cautious in interpreting the estimates obtained for the covariance matrix.

Table 5.9: Summary of center-specific cumulative log odds ratio estimates and standard errors (SE) for treatment effects with fixed and random effects heterogeneity models applied to Table 5.1. ML denotes maximum likelihood estimates from the parametric approach, while NPML denotes maximum likelihood estimates from the nonparametric approach.

			RANDOM EFFECTS		
	FIXED EFFECTS		MODEL (5.7)		
	MODEL (5.2)		ML		NPML
EFFECT	ESTIMATE	SE	ESTIMATE	SE	ESTIMATE
Center 1	3.03	0.87	2.35	0.75	2.91
Center 2	-1.62	0.95	-0.62	0.92	-0.01
Center 3	0.20	0.55	0.32	0.52	-0.04
Center 4	0.71	0.85	0.76	0.72	0.03
Center 5	2.84	0.95	2.11	0.83	2.88
Center 6	-1.06	1.21	-0.10	0.94	0.16
Center 7	1.76	0.87	1.53	0.73	1.71
Center 8	0.83	0.82	0.84	0.73	0.18

CHAPTER 6
CONCLUSIONS

6.1 Summary of Results

In this dissertation, we have developed methods for modeling longitudinal or clustered data with nominal or ordinal responses. We have concentrated on four link functions based on the logit link: the baseline-category logit link for nominal responses and the adjacent-category, continuation-ratio, and cumulative logit links for ordinal responses. To account for heterogeneity among the clustered or repeated observations, we introduced random effects linearly in the linear predictor with the fixed effects. We motivated the models from the framework of a multivariate generalized linear mixed model, yielding a general approach for modeling clustered multinomial response data. We considered both parametric and nonparametric assumptions for the distribution of the random effects. For both approaches we proposed algorithms for obtaining maximum likelihood estimates of the fixed effects parameters and the random effects distribution. We utilized a number of simulation studies to compare the two approaches, as well as to investigate inferential methods for the nonparametric approach. We then examined the use of the proposed methods for data arising from multi-center clinical trials, and concluded by proposing a score test for testing that a common association holds for all centers.

Random effects models provide a useful method for accounting for correlation among repeated or clustered observations. Through the use of the multivariate generalized linear mixed model, we have outlined a general approach for modeling clustered multinomial responses data. When the random effects are assumed to be normally distributed, one must have accurate and efficient methods for approximating the intractable integrals. We have found the adaptive Gauss-Hermite quadrature approach for approximating integrals, coupled with the direct maximization of the marginal log-likelihood to be superior to the proposed algorithms for nominal and ordinal clustered data. Specifically, we have seen that the Monte Carlo EM algorithm of Tutz and Hennevogl (1996) can provide estimates that vary quite dramatically depending on the Monte Carlo sample size that was used. Indeed, we feel that their chosen sample sizes were much too small to provide accurate estimates. In addition, the adaptive Gauss-Hermite approach is more efficient than the Gauss-Hermite approach of Hedeker and Gibbons (1994) as it provides greater accuracy for the integral approximations with fewer quadrature points. Thus one can fit models with greater numbers of random effects, while obtaining accurate approximations for the integrals. In contrast to recommendations by Hedeker and Gibbons (1994), we recommend that the number of quadrature points be increased with increasing integral dimensions, or at least set between 10 and 15 for each dimension. Under these recommendations, adaptive Gauss-Hermite quadrature is reasonable for models with up to five or six random effects.

The examples considered in this dissertation had only a small number of random effects. For high dimensional random effects, the adaptive quadrature approach will be computationally too intensive (at least with the current computing power). For such models, a Monte Carlo approach is more appropriate. We have shown that the automated Monte Carlo EM approach of Booth and Hobert (1999) can also be used for the multinomial random effects models. It is important to have good starting values that are at least close to the true maximum likelihood estimates to help reduce the computation time for the EM algorithm. The proposed pseudo-likelihood approach for the multinomial random effects models is one way of obtaining such estimates. We have seen that it can provide reasonable estimates, though it will most likely have similar small sample behavior problems as its binomial counterpart. Nonetheless, it provides a fast method for carrying out exploratory random effects modeling when the response is nominal or ordinal.

Under the assumption of normality for the random effects, we also examined the varying cumulative threshold model of Tutz and Hennevogl (1996). Though such a model allows increased flexibility in modeling a repeated ordinal response, we found it to be very unstable and, for the example considered, it provided similar results to the shifted threshold model. It is difficult to know how often such variation occurs in "real life" datasets. Indeed, the shifted threshold model produced biased estimates of the regression parameter when we simulated from a varying threshold model. However, as the correlation between thresholds neared 1.0 and the variation in the thresholds became more similar, the shifted threshold model performed adequately. Since, in practice, one would not have the software to fit the extended model, we recommend the use of the shifted threshold model as it certainly will provide better estimates than ignoring the variability completely.

An interesting alternative to the usual assumption of normality for the random effects is to estimate their distribution nonparametrically. With such an approach, one could avoid the possible misspecification of the random effects distribution. We investigated this approach and proposed an EM algorithm for obtaining the nonparametric maximum likelihood estimates of the fixed parameters and the mixing distribution in the multinomial random effects model. We found that this approach provided very similar results to the parametric approach in the examples that we considered. Simulating from a variety of random effects distributions, we found that the nonparametric approach behaved similarly to the parametric approach. In fact, the parametric approach had, in general, very little bias in the regression parameter estimate even when the true random effects distribution was far from normal. Thus, as seen by others, estimation under the assumption of normality for the random effects seems to be fairly robust to misspecification. We also examined the use of standard maximum likelihood inferential procedures for the NPML approach and found that the Wald and likelihood-ratio tests performed similarly to the corresponding tests in the parametric approach. Though the asymptotic theory does not exist for these tests when using the NPML approach, they seem to provide at least approximately correct inferences. An important issue in any mixture model is the identifiability of the model parameters. We proved a series of identifiability theorems for general mixtures of multinomial distributions, and used them to address identifiability in the multinomial NPML model. The NPML model provides a useful alternative to the parametric approach for fitting random effects models. From our work it seems that both approaches will provide similar results for most situations. As the NPML approach can converge to estimated mass points at plus and minus infinity, we recommend that one should not place too much faith in the estimated mixing distribution.

As software for random effects modeling becomes more readily available, more and more researchers will find applications for its use. We considered one such application, namely, for analyzing ordinal response data from multi-center clinical trials. If one is willing to treat the centers as random, one can model both heterogeneity in the centers and the association parameter. As such data typically has small numbers of centers, it is questionable if the heterogeneous random effects model can provide accurate results. In our simulations we found that there was a large amount of variability in the variance component estimates, especially when the number of clusters was small. However, the estimates of the association parameter were fairly accurate, as were its corresponding standard error. We also proposed score tests for testing that a common association hold for all centers. We first extended the Laplace approximated score test of Lin (1997), and then proposed an adaptive Gauss-Hermite quadrature approximated score test. Simulation studies showed that the adaptive quadrature score test performed very poorly for small to moderate center sizes. Given this, and the poor estimates for the variance components, one should be cautious in making any strong inferential statements concerning the variance components.

6.2 Future Research

There are a number of areas in this dissertation in which future research is possible. Though adaptive Gauss-Hermite quadrature seems to provide accurate approximation of integrals, and thus of the fixed and random effects parameters, there is currently not a method for evaluating the error in the approximations. Error formulas exist for Gauss-Hermite quadrature, but their evaluation requires the calculation of derivatives of extremely high order. As an alternative, one could use a similar, direct maximization of the log-likelihood approach but with a Monte Carlo technique for integration so that one could evaluate the Monte Carlo error. That is, use an approach similar to the automated EM algorithm of Booth and Hobert (1999), but with a direct maximization algorithm. The direct maximization approach with Monte Carlo integration was studied by McCulloch (1997) and was shown to perform poorly. However, we feel that he did not choose the correct candidate distribution for simulation. Since adaptive quadrature seems to work so well, one could use a multivariate normal distribution with the same mean and curvature as the integrand being approximated. This parallels exactly what is done with adaptive quadrature. One would then develop the appropriate asymptotic results to assess the Monte Carlo error in the approximations. In addition, it would be of interest to compare the adaptive quadrature and Monte Carlo direct maximization approaches and see, for example, how many samples are needed in the latter to obtain the estimates, to a certain degree of accuracy, in the former.

There are a number of open questions with regard to the NPML approach that still require answers. For one, the asymptotic theory for fixed effect tests is still unknown. Simulations suggest that the usual maximum likelihood approaches perform adequately, however, the theoretical justification is still needed. In addition, it is important to develop methods for testing the number of mass points in the distribution. In our algorithm, the number of mass points was increased until the deviance no longer changed. There may be situations, however, in which, statistically, fewer mass points are required. Such a test is under nonstandard conditions, and thus a score test might be possible. More research is also needed to completely specify the identifiability conditions for models with multiple random effects. The work of Butler and Louis (1997) has addressed some of these issues for binary random effects model. In addition, for the particular models considered here, one could modify the algorithm to allow for mass points at plus or minus infinity. It would then be interesting to compare the results of the two models.

Finally, more research is needed in studying the adaptive Gauss-Hermite approximated score test. We have seen that it performs poorly for small numbers of centers or clusters. Our simulations suggested, however, that the approximated score test would perform adequately for larger cluster sizes. Thus, one could examine other data situations, such as longitudinal studies, where the number of clusters is large. A thorough examination of the null hypothesis distribution is also needed, especially when the hypothesis implies that other covariance terms are zero.

REFERENCES

Adams, J., Wilson, M., and Wang, W. (1997), "The Multidimensional Random Coefficients Multinomial Logit Model," *Applied Psychological Measurement*, 21, 1–23.

Adams, R. J. and Wilson, M. R. (1996), "Formulating the Rasch Model as a Mixed Coefficients Multinomial Logit," In Engelhard, G. and Wilson, M. R., editors, *Objective Measurement: Theory and Practice*, volume 3, pages 143–166, Norwood, NJ: Ablex.

Agresti, A. (1990), *Categorical Data Analysis*, New York: Wiley.

Agresti, A. (1993a), "Computing Conditional Maximum Likelihood Estimates for Generalized Rasch Models Using Simple Loglinear Models with Diagonal Parameters," *Scandinavian Journal of Statistics*, 20, 63–71.

Agresti, A. (1993b), "Distribution-free Fitting of Logit Models With Random Effects of Repeated Categorical Responses," *Statistics in Medicine*, 12, 1969–1987.

Agresti, A. and Hartzel, J. (1999), "Tutorial in Biostatistics: Strategies for Comparing Treatments on a Binary Response with Multi-Center Data," *Statistics in Medicine*, in press.

Agresti, A. and Lang, J. B. (1993), "A Proportional Odds Models with Subject-specific Effects for Repeated Ordered Categorical Responses," *Biometrika*, 80, 527–534.

Aitkin, M. (1996), "A General Maximum Likelihood Analysis of Overdispersion in Generalized Linear Models," *Statistics and Computing*, 6, 251–262.

Aitkin, M. (1999), "A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models," *Biometrics*, 55, 117–128.

Aitkin, M. and Aitkin, I. (1995), "A Hybrid EM/Gauss-Newton Algorithm for Maximizing Likelihood in Mixture Distributions," *Statistics and Computing*, 6, 127–130.

Aitkin, M. and Alfo, M. (1998), "Regression Models for Binary Longitudinal Responses," *Statistics and Computing*, 8, 289–307.

Andersen, E. B. (1973), "Conditional Inference for Multiple-Choice Questionnaires," *British Journal of Mathematical and Statistical Psychology*, 26, 31–44.

Andersen, E. B. (1980), *Discrete Statistical Models With Social Science Applications*, Amsterdam: North-Holland/Elsevier.

Anderson, D. A. and Aitkin, M. (1985), "Variance Component Models with Binary Response: Interviewer Variability," *J. Roy. Statis. Soc. B*, 47, 203–210.

Bartholomew, D. (1987), *Latent Variable Models and Factor Analysis*, New York: Oxford Press.

Ben-Akiva, M. and Lerman, S. (1985), *Discrete Choice Analysis*, Cambridge: The MIT Press.

Bohning, D. (1995), "A Review of Reliable Maximum Likelihood Algorithms for Semiparametric Mixture Models," *Journal of Statistical Planning and Inference*, 47, 5–28.

Bohning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. (1994), "The Distribution of the Likelihood Ratio for Mixtures of Densities from the One-Parameter Exponential Family," *Annals of the Institute of Statistical Mathematics*, 46, 373–388.

Booth, J. G. and Hobert, J. P. (1998), "Standard Errors of Prediction in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 93, 262–272.

Booth, J. G. and Hobert, J. P. (1999), "Maximizing Generalized Linear Mixed Model Likelihoods with an automated Monte Carlo EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 61, 265–285.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25.

Breslow, N. E. and Lin, X. (1995), "Bias Correction in Generalised Linear Mixed Models With a Single Component of Dispersion," *Biometrika*, 82, 81–91.

Broyden, C. (1970), "The Convergence of a Class of Double-rank Minimization Algorithms," *Journal of the Institute of Mathematics and its Applications*, 6, 76–90.

Bryk, A. and Raudenbush, A. (1992), *Hierarchical Linear Models*, Thousand Oaks, California: Sage Publications, Inc.

Butler, S. M. and Louis, T. A. (1992), "Random Effects Models with Nonparametric Priors," *Statistics in Medicine*, 11, 1981–2000.

Butler, S. M. and Louis, T. A. (1997), "Consistency of Maximum Likelihood Estimators in General Random Effects Models for Binary Data," *The Annals of Statistics*, 25, 351–377.

Chan, J. S. K. and Kuk, A. Y. C. (1997), "Maximum Likelihood Estimation for Probit-Linear Mixed Models with Correlated Random Effects," *Biometrics*, 53, 86–97.

Chant, D. (1974), "On Asymptotic Tests of Composite Hypotheses in Nonstandard Conditions," *Biometrika*, 61, 291–298.

Clogg, C. (1979), "Some Latent Structure Models for the Analysis of Likert-type Data," *Social Science Research*, 8, 287–301.

Collett, D. (1991), *Modelling Binary Data*, London: Chapman and Hall.

Conaway, M. (1989), "Analysis of Repeated Categorical Measurements with Conditional Likelihood Methods," *Journal of the American Statistical Association*, 84, 53–62.

Conaway, M. R. (1990), "A Random Effects Model for Binary Data," *Biometrics*, 46, 317–328.

Coull, B. and Agresti, A. (2000), "Random Effects Modeling of Multiple Binary Responses Using the Multivariate Binomial Logit-Normal Distribution," *Biometrics*, 56, 162–168.

Cox, D. R. (1972), "Regression Models and Life-Tables (with Discussion)," *Journal of the Royal Statistical Society Series B*, 34, 187–220.

Davies, R. (1993), "Nonparametric Control for Residual Heterogeneity in Modelling Recurrent Behavior," *Computational Statistics & Data Analysis*, 16, 143–160.

Davies, R. and Pickles, A. (1987), "A Joint Trip Timing, Store Choice Model Including Feedback Effects and Nonparametric Control for Omitted Variables," *Transportation Research A*, 21, 345–361.

Davies, R. B. (1987), "Mass Point Methods for Dealing with Nuisance Parameters in Longitudinal Studies," In Crouchley, R., editor, *Longitudinal Data Analysis*, pages 88–109, Hants, England: Avebury, Aldershot.

Dempster, A. P., Laird, N. M., and Rubin, D. A. (1977), "Maximum Likelihood Estimation From Incomplete Data Via the EM Algorithm (with Discussion)," *Journal of the Royal Statistical Society Series B*, 39, 1–38.

Dietz, E. and Bohning, D. (1995), "Statistical Inference Based on a General Model of Unobserved Heterogeneity," In Steckel-Berger, G., Francis, B., Hatzinger, R., and Seeber, G., editors, *Lecture Notes in Statistics: Statistical Modelling*, volume 104, pages 75–82, New York: Springer-Verlag.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford: Clarendon Press.

Doornick, J. A. (1998), *Object-Oriented Matrix Programming Using Ox 2.0*, Kent, England: Timberlake Consultants, Ltd.

Efron, B. and Hinkley, D. V. (1978), "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information (C/R: p482-487)," *Biometrika*, 65, 457–481.

Engel, B. (1998), "A Simple Illustration of the Failure of PQL, IRREML and APHL as Approximate ML Methods for Mixed Models for Binary Data," *Biometrical Journal*, 40, 141–154.

Engel, B. and Keen, A. (1994), "A Simple Approach for the Analysis of Generalized Linear Mixed Models," *Statistica Neerlandica*, 48, 1–22.

Ezzet, F. and Whitehead, J. (1991), "A Random Effects Model For Ordinal Responses From A Crossover Trial," *Statistics in Medicine*, 10, 901–907.

Fahrmeir, L. and Tutz, G. (1994), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer-Verlag New York, Inc.

Fienberg, S. E. (1986), "The Rasch Model," In *Encyclopedia of Statistical Sciences*, volume 7, pages 627–632, New York: Wiley.

Fleiss, J. (1986), "Analysis of Data From Multiclinic Trials," *Controlled Clinical Trials*, 10, 237–243.

Fletcher, R. (1970), "A New Approach to Variable Metric Algorithms," *Computer Journal*, 13, 317–322.

Follmann, D. A. and Lambert, D. (1989), "Generalizing Logistic Regression By Nonparametric Mixing," *Journal of the American Statistical Association*, 84, 295–300.

Follmann, D. A. and Lambert, D. (1991), "Identifiability of Finite Mixtures of Logistic Regression Models," *Journal of Statistical Planning and Inference*, 27, 375–381.

Geweke, J. (1996), *Handbook of Computational Statistics*, chapter 15, New York: Elsevier.

Gilmour, A. R., Anderson, R. D., and Rae, A. L. (1985), "The Analysis of Data by a Generalized Linear Mixed Model," *Biometrika*, 72, 593–599.

Goldfarb, D. (1970), "A Family of Variable Metric Methods Derived by Variational Means," *Mathematics of Computation*, 24, 23–26.

Golub, G. H. (1973), "Some Modified Matrix Eigenvalue Problems," *SIAM Review*, 15, 318–334.

Graham, A. (1981), *Kronecker Products and Matrix Calculus with Applications*, Horwood, New York: Halsted Press.

Grizzle, J. E. (1987), "Letter to the Editor," *Controlled Clinical Trials*, 8, 392–393.

Harville, D. A. (1977), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems (with Discussion)," *Journal of the American Statistical Association*, 72, 320–385.

Harville, D. A. and Mee, R. W. (1984), "A Mixed-model Procedure for Analyzing Ordered Categorical Data," *Biometrics*, 40, 393–408.

Heagerty, P. J. (1999), "Marginally Specified Logistic-Normal Models for Longitudinal Binary Data," *Biometrics*, 55, 688–698.

Heckman, J. J. and Singer, B. (1984), "A Method For Minimizing the Impact of Distributional Assumptions in Econometric Models of Duration," *Econometrica*, 52, 271–320.

Hedeker, D. (2000), "MIXNO: A Computer Program for Mixed-Effects Nominal Logistic Regression," *Computer Methods and Programs in Biomedicine*, in press.

Hedeker, D. and Gibbons, R. D. (1994), "A Random-effects Ordinal Regression Model for Multilevel Analysis," *Biometrics*, 50, 933–944.

Henderson, C. R. (1975), "Best Linear Unbiased Estimation and Prediction Under a Selection Model," *Biometrics*, 31, 423–447.

Hinde, J. P. (1982), "Compound Regression Models," In Gilchrist, R., editor, *GLIM 82: International Conference for Generalized Linear Models*, pages 109–121, New York: Springer.

Hinde, J. P. and Demetrio, C. G. B. (1998), "Overdispersion: Models and Estimation," *Computational Statistics & Data Analysis*, 27, 151–170.

Hobert, J. and Casella, G. (1996), "The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models," *Journal of the American Statistical Association*, 91, 1461–1473.

Jacqmin-Gadda, H. and Commenges, D. (1995), "Tests of Homogeneity for Generalized Linear Models," *Journal of the American Statistical Association*, 90, 1237–1246.

Jansen, J. (1990), "On the Statistical Analysis of Ordinal Data When Extravariation is Present," *Applied Statistics*, 39, 74–85.

Johnson, M. (1987), *Multivariate Statistical Simulation*, New York: John Wiley & Sons, Inc.

Jones, B., Teather, D., Wang, J., and Lewis, J. (1998), "A Comparison of Various Estimators of a Treatment Difference for a Multi-centre Clinical Trial," *Statistics in Medicine*, 17, 1767–1777.

Jones, R. H. (1993), *Longitudinal Data with Serial Correlation: A State-Space Approach*, London: Chapman and Hall.

Karim, M. R. and Zeger, S. L. (1992), "Generalized Linear Models With Random Effects; Salamander Mating Revisited," *Biometrics*, 48, 631–644.

Kaufmann, H. (1988), "On Existence and Uniqueness of Maximum Likelihood Estimates in Quantal and Ordinal Response Models," *Metrika*, 35, 291–313.

Keen, A. and Engel, B. (1997), "Analysis of a Mixed Model for Ordinal Data by Iterative Re-Weighted REML," *Statistica Neerlandica*, 51, 129–144.

Kiefer, J. and Wolfowitz, J. (1972), "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Nuisance Parameters," *Annals of Mathematical Statistics*, 27, 887–906.

Laird, N. (1978), "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of the American Statistical Association*, 73, 805–811.

Laird, N. M. and Ware, J. H. (1982), "Random Effects Models for Longitudinal Data," *Biometrics*, 38, 963–974.

Lesperance, M. and Kalbfleisch, J. D. (1992), "An Algorithm for Computing the Nonparametric MLE of a Mixing Distribution," *Journal of the American Statistical Association*, 87, 120–126.

Liang, K. Y. (1987), "A Locally Most Powerful Test for Homogeneity with Many Strata," *Biometrika*, 74, 259–264.

Liang, K. Y. and Zeger, S. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lin, X. (1997), "Variance Component Testing in Generalised Linear Models with Random Effects," *Biometrika*, 84, 309–326.

Lin, X. and Breslow, N. E. (1996), "Bias Correction in Generalised Linear Mixed Models With Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007–1016.

Lindsay, B. (1983a), "The Geometry of Mixture Likelihoods: A General Theory," *The Annals of Statistics*, 11, 86–94.

Lindsay, B. (1983b), "The Geometry of Mixture Likelihoods, Part II: The Exponential Family," *The Annals of Statistics*, 11, 783–792.

Lindsay, B. (1989), "Moment Matrices: Applications in Mixtures," *The Annals of Statistics*, 17, 722–740.

Lindsay, B., Clogg, C. C., and Grego, J. (1991), "Semiparametric Estimation in the Rasch Model and Related Exponential Response Models, Including a Simple Latent Class Model for Item Analysis," *Journal of the American Statistical Association*, 86, 96–107.

Lindsey, J., Jones, B., and Ebbutt, A. (1997), "Simple Models for Repeated Ordinal Responses with an Application to a Seasonal Rhinitis Clinical Trial," *Statistics in Medicine*, 16, 2873–2882.

Lindsey, J. K. (1993), *Models for Repeated Measurements*, Oxford: Oxford University Press.

Lipsitz, S. R., Kim, K., and Zhao, L. (1994), "Analysis of Repeated Categorical Data Using Generalized Estimating Equations," *Statistics in Medicine*, 13, 1149–1163.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991), "Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as a Measure of Association," *Biometrika*, 78, 153–160.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996), *SAS System for Mixed Models*, Cary, NC: SAS Institute, Inc.

Liu, C. and Rubin, D. (1994), "The ECME Algorithm: A Simple Extension of EM and ECM with Faster Monotone Convergence," *Biometrika*, 81, 633–648.

Liu, C. and Sun, D. (1997), "Accerlation of EM Algorithm for Mixture Models Using ECME," In *Proceedings of the Statistical Computing Section*, pages 109–114, Washington: The American Statistical Association.

Liu, I. and Agresti, A. (1996), "Mantel-Haenszel-Type Inference for Cumulative Odds Ratios with a Stratified Response," *Biometrics*, 52, 1223–1234.

Liu, Q. and Pierce, D. A. (1994), "A Note on Gauss-Hermite Quadrature," *Biometrika*, 81, 624–629.

Longford, N. T. (1993), *Random Coefficient Models*, New York: Oxford University Press.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 44, 226–233.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press.

Mantel, N. and Haenszel, W. (1959), "Statistical aspects of the analysis of data from retrospective studies of disease," *Journal of the National Cancer Institute*, 22, 719–748.

Masters, G. (1985), "A Comparison of Latent Trait and Latent Class Analyses of Likert-type Data," *Psychometrika*, 50, 69–82.

McCullagh, P. (1980), "Regression Models for Ordinal Data (With Discussion)," *Journal of the Royal Statistical Society, Series B*, 42, 109–142.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman and Hall.

McCulloch, C. E. (1994), "Maximum Likelihood Variance Components Estimation for Binary Data," *Journal of the American Statistical Association*, 89, 330–335.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Meng, X. and Rubin, D. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–279.

Natarajan, R. and McCulloch, C. E. (1995), "A Note on the Existence of the Posterior Distribution for a Class of Mixed Models for Binomial Responses," *Biometrika*, 82, 639–643.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. D. (1992), "The Effects of Mixture Distribution Misspecification When Fitting Mixed-effects Logistic Models," *Biometrika*, 79, 755–762.

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1991), "A Comparison of Cluster-specific and Population-averaged Approaches for Analyzing Correlated Binary Data," *International Statistical Review*, 59, 25–35.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R. (1996), "A Survey of Methods for Analyzing Clustered Binary Response Data," *International Statistical Review*, 64, 89–118.

Pierce, D. A. and Sands, B. R. (1975), "Extra-Bernoulli Variation in Regression of Binary Data," Technical Report 46, Oregon State University, Dept. of Statistics, Corvallis.

Pinheiro, J. C. and Bates, D. M. (1995), "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-effects Model," *Journal or Computational and Graphical Statistics*, 4, 12–35.

Prakasa Rao, B. (1987), *Asymptotic Theory of Statistical Inference*, New York: Wiley.

Prakasa Rao, B. (1992), *Identifiability in Stochastic Models: Characterization of Probability Distributions*, New York: Academic Press, Inc.

Prentice, R. L. (1988), "Correlated Binary Regression with Covariates Specific to Each Binary Observation," *Biometrics*, 44, 1033–1084.

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985), "The Developmental Toxicity of Ethylene Glycol in Rats and Mice," *Toxicology and Applied Pharmacology*, 81, 113–127.

Randall, J. (1989), "The Analysis of Sensory Data by Generalized Linear Models," *Biometrical Journal*, 31, 781–793.

Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, New York: Wiley, 2nd edition.

Rasch, G. (1961), "On General Laws and the Meaning of Measurement in Psychology," In Neyman, J., editor, *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Vol 4*, Berkeley: University of California Press.

Redner, R. and Walker, H. (1984), "Mixture Densities, Maximum Likelihood, and the EM Algorithm," *Society for Industrial and Applied Mathematics*, 26, 195–239.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: John Wiley & Sons, Inc.

Self, S. and Liang, K.-Y. (1987), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, 82, 605–610.

Senn, S. (1998), "Some Controversies in Planning and Analysing Multi-centre Trials," *Statistics in Medicine*, 17, 1753–1765.

Shanno, D. (1970), "Conditioning of Quasi-Newton Methods for Function Minimization," *Mathematics of Computation*, 24, 647–657.

Skene, A. M. and Wakefield, J. C. (1990), "Hierarchical Models for Multicentre Binary Response Studies," *Statistics in Medicine*, 9, 919–929.

Stiratelli, R., Laird, N. M., and Ware, J. H. (1984), "Random Effects Models for Serial Observations with Binary Responses," *Biometrics*, 40, 961–971.

Stroud, A. and Secrest, D. (1966), *Gaussian Quadrature Formulas*, Englewood Cliffs, New Jersey: Prentice Hall.

Swallow, W. and Monahan, J. (1984), "Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components," *Technometrics*, 28, 47–57.

Tanner, M. A. (1993), *Tools for Statistical Inference: Observed Data and Data Augmentation (2nd ed.)*, Berlin: Springer-Verlag.

Tanner, M. A. (1996), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Berlin: Springer-Verlag.

Teicher, H. (1963), "Indentifiability of Finite Mixtures," *Annals of Mathematical Statistics*, 34, 1265–1269.

Teicher, H. (1967), "Indentifiability of Mixtures of Product Measures," *Annals of Mathematical Statistics*, 38, 1300–1302.

Ten Have, T. R. (1996), "A Mixed Effects Model for Multivariate Ordinal Response Data Including Correlated Discrete Failure Times with Ordinal Responses," *Biometrics*, 52, 473–491.

Ten Have, T. R., Landis, J. R., and Hartzel, J. (1996), "Population-Averaged and Cluster-Specific Models for Clustered Ordinal Response Data," *Statistics in Medicine*, 15, 2573–2588.

Ten Have, T. R. and Uttal, D. H. (1994), "Subject-Specific and Population-Averaged Continuation Ratio Logit Models for Multiple Discrete Time Survival Profiles," *Applied Statistics*, 32, 371–384.

Thurstone, L. (1927), "A Law of Comparitive Judgement," *Psychological Review*, 34, 273–286.

Titterington, D., Smith, A., and Makov, U. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Tjur, T. (1982), "A Connection Between Rasch's Item Analysis Model and a Multiplicative Poisson Model," *Scandinavian Journal of Statistics*, 9, 23–30.

Tutz, G. and Hennevogl, W. (1996), "Random Effects in Ordinal Regression Models," *Computational Statistics & Data Analysis*, 22, 537–557.

Uesaka, H. (1993), "Test for Interaction Between Treatment and Stratum with Ordinal Responses," *Biometrics*, 49, 123–129.

Wedderburn, R. W. M. (1974), "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, 61, 439–447.

Wedel, M. and DeSarbo, W. S. (1995), "A Mixture Likelihood Approach for Generalized Linear Models," *Journal of Classification*, 12, 21–55.

Williams, D. A. (1982), "Extra-Binomial Variation in Logistic Linear Models," *Applied Statistics*, 31, 144–148.

Wolfinger, R. and O'Connell, M. (1993), "Generalized Linear Mixed Models: A Pseudo-likelihood Approach," *Journal of Statistical Computation and Simulation*, 48, 233–243.

Wolfram, S. (1996), *The Mathematica Book (3rd ed.)*, New York: Wolfram Media/Cambridge University Press.

Wood, A. and Hinde, J. (1987), "Binomial Variance Component Models with a Nonparametric Assumption Concerning Random Effects," In Crouchley, R., editor, *Longitudinal Data Analysis*, pages 110–128, Hants, England: Avebury, Aldershot.

Zeger, S. and Liang, K. Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics*, 42, 121–130.

Zeger, S. L. and Karim, M. R. (1991), "Generalized Linear Models With Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988), "Models for Longitudinal Data: A Generalized Estimating Equation Approach (Corr: V45 P347)," *Biometrics*, 44, 1049–1060.

BIOGRAPHICAL SKETCH

Jonathan Seth Hartzel was born February 13, 1971, in Lansdale, Pennsylvania, along with his twin sister Kristin, to Norm and Judy Hartzel, and their older brother Nathan. Soon after birth, Jonathan moved to Telford, Pennsylvania, where he lived until August of 1989 when he left for college. Both Jonathan and Kristin attended Messiah College in Grantham, Pennsylvania, where Jonathan played collegiate soccer and graduated with a Bachelor of Arts degree in mathematics in June of 1993. He then moved to Elizabethtown, Pennsylvania, where he worked for the Center for Biostatistics and Epidemiology in the College of Medicine of Pennsylvania State University. During his time there, he met his future wife, Tracy Ann Plieninger. In August of 1994, Jonathan moved to Gainesville, Florida, and began graduate school in the Department of Statistics at the University of Florida.

In his first three years in the Department of Statistics, Jonathan worked in the Biostatistics Consulting Lab, where he provided consulting and statistical support for doctors and medical students in Shands Medical Center. During his final years in the department, Jonathan has worked under the direction of his advisor, Dr. Alan Agresti, as a research assistant. In August of 1995, Jonathan married Tracy in Oreland, Pennsylvania, following which Tracy joined Jonathan in Gainesville, taking employment with the Department of Biostatistics at the University of Florida. Jonathan received his Master of Statistics degree in December of 1996 and plans to receive his Ph.D. in December of 1999. Jonathan and Tracy look forward to moving to Blue Bell, Pennsylvania, with their two Rhodesian Ridgebacks, Riley and Kendi, where Jonathan has accepted a position with Merck Research Laboratories, and Tracy has accepted a position with Wyeth-Ayerst pharmaceutical company.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Alan Agresti

Alan G. Agresti, Chairman
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Malay Ghosh

Malay Ghosh
Distinguished Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

James P. Hobert
Assistant Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Ramon C. Littell
Professor of Statistics

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a dissertation for the degree of Doctor of Philosophy.

Gary Miller
Associate Professor of Mechanical
Engineering

This dissertation was submitted to the Graduate Faculty of the Department of Statistics in the College of Liberal Arts and Sciences and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

December 1999

Dean, Graduate School